Towards Understanding the Trade-off Between Accuracy and Adversarial Robustness

Congyue Deng[*1], **Yi Tian**[*1]
dengcy16@mails.tsinghua.edu.cn, tianyi15@mails.tsinghua.edu.cn
[1]Tsinghua University, Beijing, China

1 Introduction

We aim to promote understanding of the phenomenon of adversarial examples by analyzing a trade-off between accuracy and adversarial robustness in an idealized setting and under the infinite data assumption.

- The classifier with the highest standard accuracy provably differs from that with the highest adversarial robustness, which is obtainable by adversarial training. Between the standard optimal classifier and the adversarially optimal classifier, we can find classifiers that are optimal in the sense of linear combinations of these two goals.
- The distance between the standard and the adversarially optimal decision hyperplanes can be both lower and upper bounded, and both bounds are proportional to the attack radius ε. Specifically, under ℓ_∞-attack the distance is $\Theta(\sqrt{d}\varepsilon)$ with d the dimensionality.
- Different training strategies, including standard training, adversarial training, and data-randomized training favor accuracy and adversarial robustness differently. The in-between classifiers that balance the trade-off can be obtained by data-randomized training with different randomization parameters.
- For some data distributions, it is possible to improve the adversarial robustness of a classifier significantly at the price of a slight accuracy decrease.

2 Our Setting and the Definitions

We consider a simple yet useful setting: binary classification over two spherical Gaussian distributions.

Definition (Standard Error). Let $\mathcal{P}$ be a distribution on $\mathbb{R}^d \times \{\pm 1\}$. Then the classification error for a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\beta := \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{P}}[f(\boldsymbol{x}) \neq y]$.

Definition ($\mathcal{B}_p^\varepsilon$-Robust Classification Error). Let $\mathcal{P}$ be a distribution on $\mathbb{R} \times \{\pm 1\}$. For $\boldsymbol{x} \in \mathbb{R}^d$, we denote the ε-neighborhood under ℓ_p-distance by $\mathcal{B}_p^\varepsilon(\boldsymbol{x}) = \left\{\boldsymbol{x}' \in \mathbb{R}^d \middle| \|\boldsymbol{x}' - \boldsymbol{x}\|_p < \varepsilon\right\}$. Then the $\mathcal{B}_p^\varepsilon$-robust classification error for a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is defined as $\widetilde{\beta} := \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{P}}\left[\exists \boldsymbol{x}' \in \mathcal{B}_p^\varepsilon(\boldsymbol{x}) : f(\boldsymbol{x}') \neq y\right]$.

3 A Trade-off between Accuracy and Adversarial Robustness

Accuracy and Adversarial Robustness Lead to Different Optimal Classifiers

Theorem 1. Consider the linear classification task on the data set with two classes following d-dimensional Gaussian distributions $\mathcal{N}_d(\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I})$ and $\mathcal{N}_d(\boldsymbol{\mu}_2, \sigma_2^2\boldsymbol{I})$ respectively, with $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. If $\sigma_1 \neq \sigma_2$, then the standard optimal decision hyperplane $W^\star$ and the $\mathcal{B}_p^\varepsilon$-robust optimal decision hyperplane $\widetilde{W}^\star$ are two **different** parallel hyperplanes in $\mathbb{R}^d$. Moreover, for any $\lambda \in (0, +\infty)$,

$$W_\lambda^\star = \arg\min_W \beta(W) + \lambda\widetilde{\beta}(W)$$

yields a decision hyperplane lying between $W^\star$ and $\widetilde{W}^\star$ and parallel to them.

Accuracy and Adversarial Robustness Lead to Different Optimal Classifiers

Theorem 2. Let $\boldsymbol{e}_{12,\infty} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_\infty$ be the ℓ_∞ norm unit vector in the direction of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Suppose $\boldsymbol{e}_{12,\infty}$ is uniformly distributed on an ℓ_∞ unit sphere. Then under ℓ_∞-attack, the average distance between the standard optimal decision hyperplane $W^\star$ and the $\mathcal{B}_\infty^\varepsilon$-robust optimal decision hyperplane $\widetilde{W}^\star$ has a lower bound

$$\mathbb{E}_{\boldsymbol{e}_{12,\infty}}\left\|\widetilde{W}^* - W^*\right\|_2 \geqslant \frac{\sigma_2 - \sigma_1}{\sigma_2 + \sigma_1}\frac{d+2}{3\sqrt{d}}\varepsilon.$$

Theorem 3. Under ℓ_∞-attack, the distance between the two optimal decision hyperplanes has an upper bound

$$\left\|\widetilde{W}^* - W^*\right\|_2 \leqslant \frac{\sigma_2^2 + \sigma_1^2}{\sigma_2^2 - \sigma_1^2}\sqrt{d}\varepsilon.$$

4 Balancing the Trade-off

Adversarial Training

The adversarial training that we consider is to replace $\boldsymbol{x}$ by $\boldsymbol{x}' \in \mathcal{B}_p^\varepsilon(\boldsymbol{x})$ as

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}} \max_{\boldsymbol{x}' \in \mathcal{B}_p^\varepsilon} \ell(\theta; \boldsymbol{x}').$$

$\mathcal{B}_p^\varepsilon$ adversarial training results in the $\mathcal{B}_p^\varepsilon$-robust optimal decision hyperplane $\widetilde{W}^\star$ when the amount of data approaches infinity. The adversarial training, as defined above, corresponds to $\lambda = \infty$, and leads to the most robust classifier given infinite data.

Data-Randomized Training

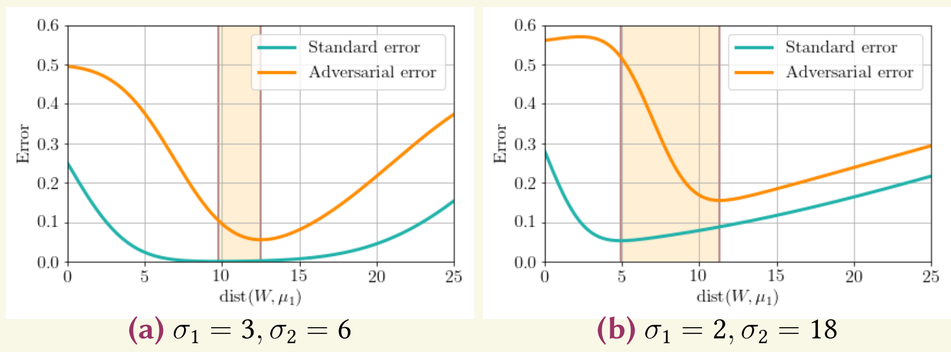
Data-randomized training refers to replacing $\boldsymbol{x}$ by $\boldsymbol{x} + \boldsymbol{\delta}$ during training as

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}} \ell(\theta; \boldsymbol{x} + \boldsymbol{\delta}), \tag{1}$$

where $\boldsymbol{\delta}$ is a random perturbation drawn from a certain distribution. We show that for data-randomized training, λ is controlled by the randomization parameter δ_σ, and can take any value between $[0, +\infty]$ as δ_σ varies.

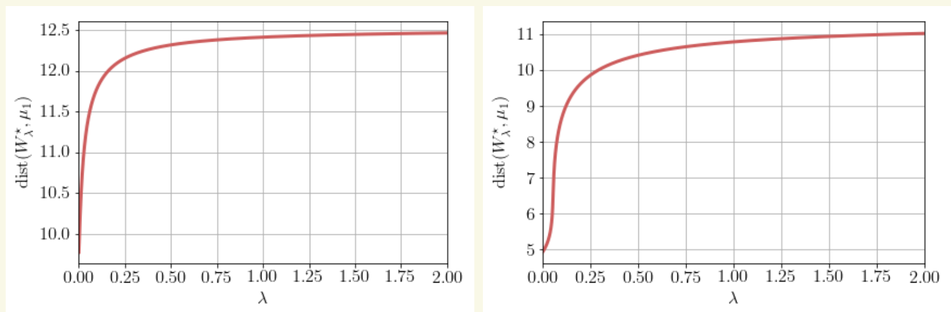
5 Numerical Case Studies

Between the two minimum points, the curve of standard error declines gradually while the curve of adversarial error declines steeply. This is a demonstration of the cases where adversarial robustness can be improved significantly at the price of a slight accuracy decrease.



(a) $\sigma_1 = 3, \sigma_2 = 6$ **(b)** $\sigma_1 = 2, \sigma_2 = 18$

The change of $W_\lambda^\star$ becomes slow after λ exceeds a certain value. Note that in these cases this value is quite small (approximately 0.25), that is, even for an objective assigning more weight on accuracy than robustness, it still yields a decision hyperplane that is very close to the adversarially optimal decision hyperplane $\widetilde{W}^\star$.



(a) $\sigma_1 = 3, \sigma_2 = 6$ **(b)** $\sigma_1 = 2, \sigma_2 = 18$