Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

Chengyu Dong[1], Liyuan Liu[2], Jingbo Shang[1]

[1] University of California, San Diego

[2] Microsoft Research

Our findings and contributions

- **Label noise implicitly exists in adversarial training**
- Robust overfitting is in fact an early part of a double descent, and can be explained by such label noise
- Alternative labeling of adversarial examples can reduce the label noise and mitigate robust overfitting

Background: Adversarial training

- **Adversarial training is one of the most effective ways to enhance the adversarial robustness of deep neural networks**

Step 1: Generate a set of adversarial examples by perturbing their clean counterparts

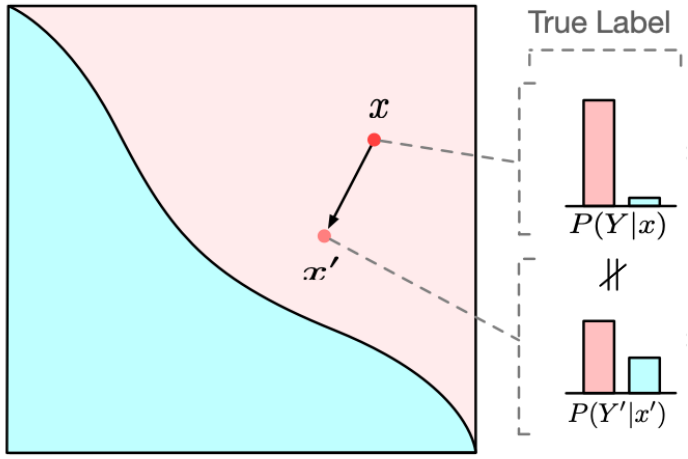
$$x' = \arg\max_{z \in \mathcal{B}_\varepsilon(x)} \ell(f(z), y)$$

Step 2: Train the classifier on the set of adversarial examples

$$\theta^* = \arg\min_{\theta} \frac{1}{|\mathcal{D}'|} \sum_{(x', \tilde{y}') \in D'} \ell(f_\theta(x'), \tilde{y}').$$

Label noise implicitly exists in adversarial training

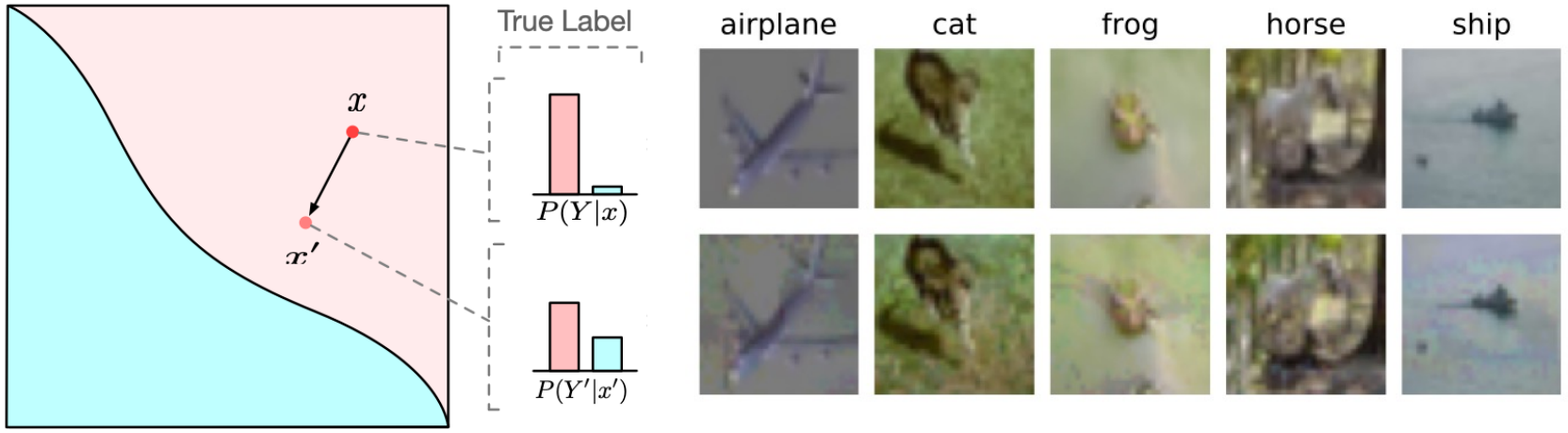
- **Adversarial training practice introduces distribution mismatch between <u>true label distribution</u> and <u>assigned label distribution</u> of adversarial examples**



-> Adversarial perturbation can distort the data semantics, which means adversarial examples have different true label distribution from their clean counterparts

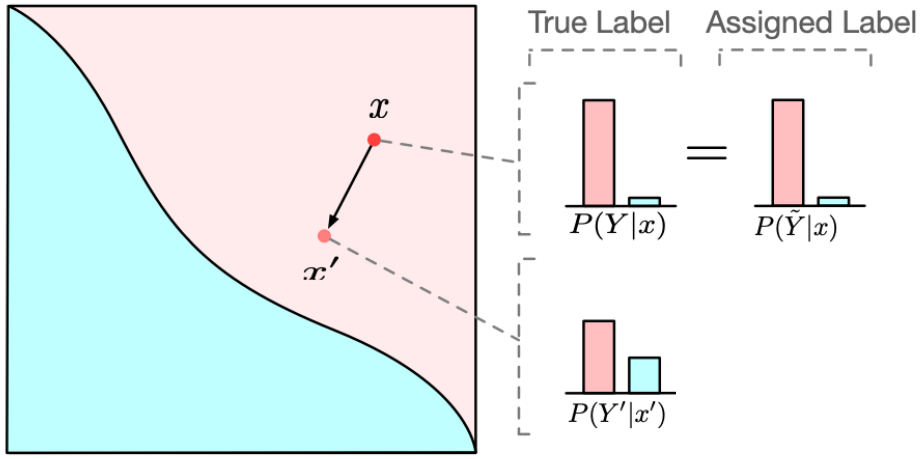
Label noise implicitly exists in adversarial training

- **Adversarial training practice introduces distribution mismatch between <u>true label distribution</u> and <u>assigned label distribution</u> of adversarial examples**



Label noise implicitly exists in adversarial training

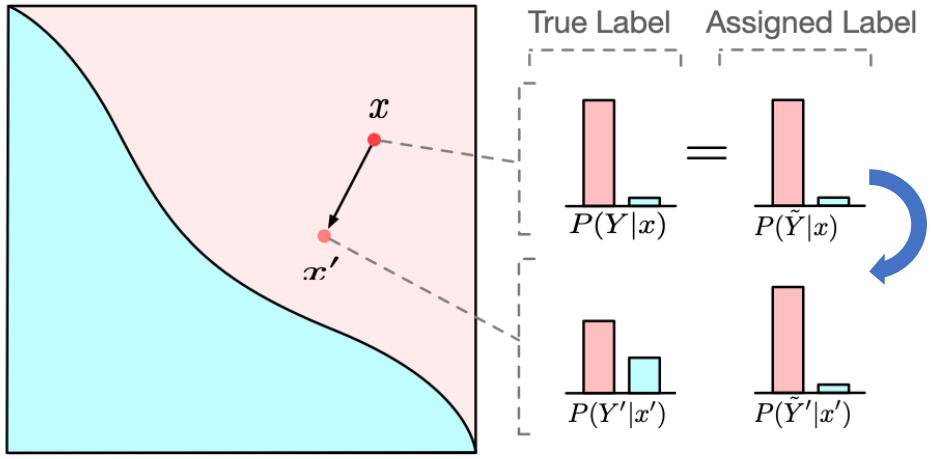
- **Adversarial training practice introduces distribution mismatch between <u>true label distribution</u> and <u>assigned label distribution</u> of adversarial examples**



-> Assume no label noise in the clean dataset

Label noise implicitly exists in adversarial training

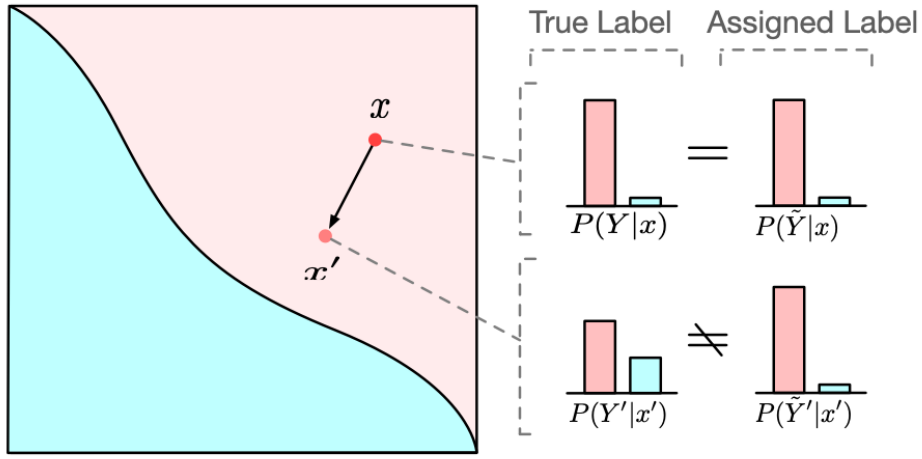
- **Adversarial training practice introduces distribution mismatch between <u>true label distribution</u> and <u>assigned label distribution</u> of adversarial examples**



-> Assigned labels of adversarial examples are directly copied from that of their clean counterparts

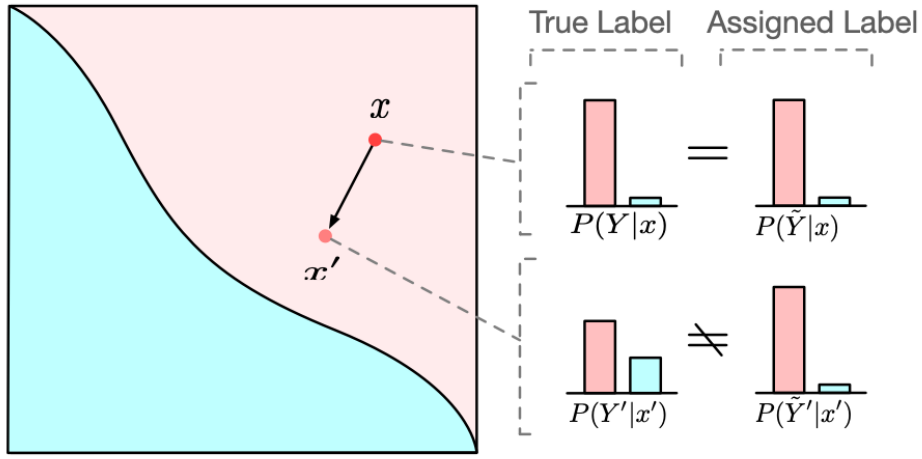
Label noise implicitly exists in adversarial training

- **Adversarial training practice introduces distribution mismatch between <u>true label distribution</u> and <u>assigned label distribution</u> of adversarial examples**



Label noise implicitly exists in adversarial training

- Adversarial training practice introduces distribution mismatch between true label distribution and assigned label distribution of adversarial examples



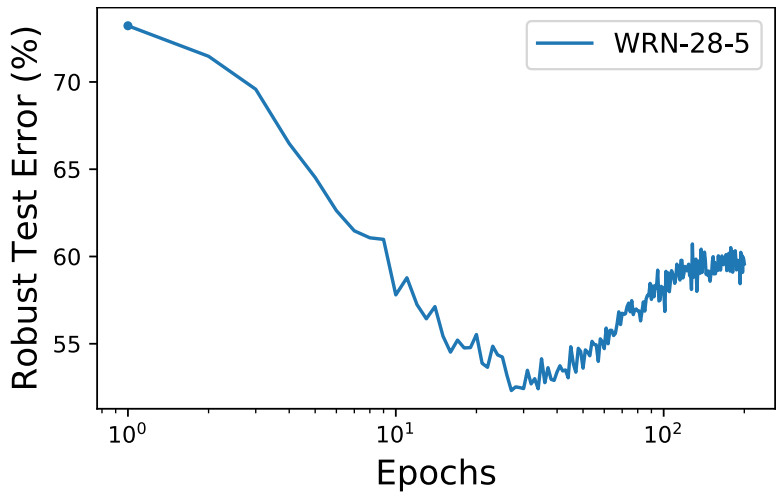
- **Distribution mismatch causes label noise**

Coupling inequality: $P(Y' \neq \tilde{Y}'|x') \geq \|P(Y'|x') - P(\tilde{Y}'|x')\|_{\mathrm{TV}}$

Label noise *Distribution mismatch*

Effect of Label noise in adversarial training

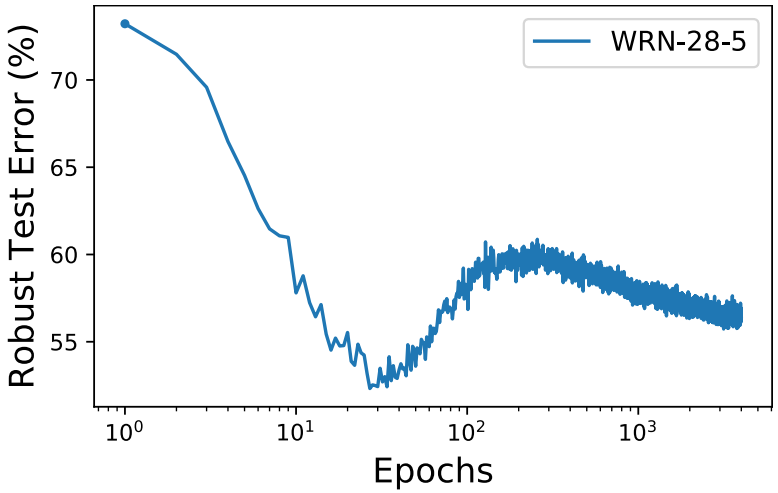
- **Label noise in adversarial training explains robust overfitting**



Robust overfitting

Effect of Label noise in adversarial training

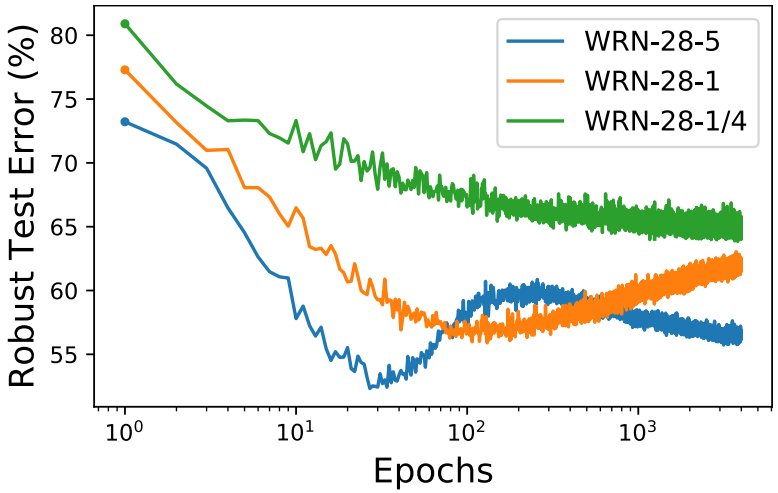
- Label noise in adversarial training explains robust overfitting
- **Robust overfitting should be double descent based on this understanding**



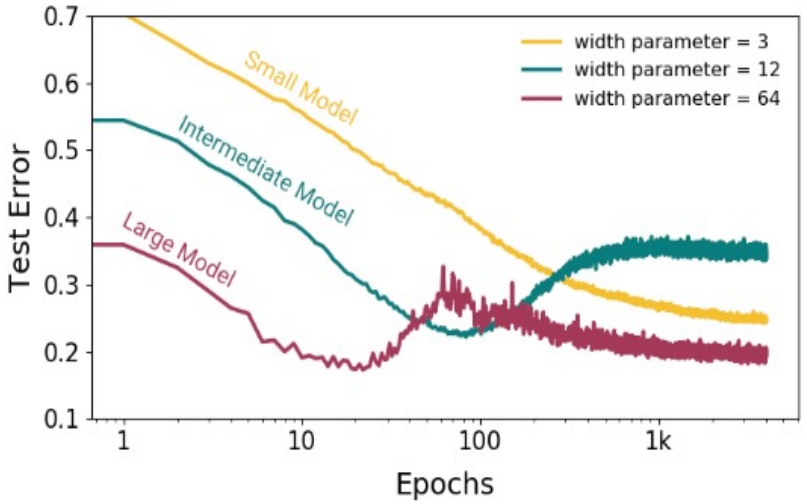
Double descent in adversarial training

Effect of Label noise in adversarial training

- Label noise in adversarial training explains robust overfitting
- **Robust overfitting should be double descent based on this understanding**



Double descent in adversarial training



Double descent in standard training (Nakkiran et al., 2020)

Mitigate label noise in adversarial training

- **Label noise in adversarial training comes from the mismatch between the true label distribution and assigned label distribution**

$$P(Y' \neq \tilde{Y}'|x') \geq \|P(Y'|x') - P(\tilde{Y}'|x')\|_{\mathrm{TV}}$$

Mitigate label noise in adversarial training

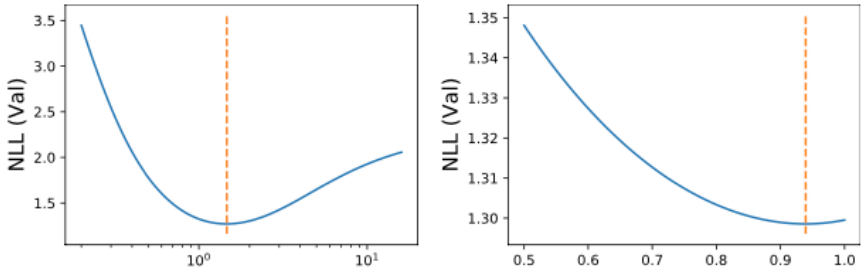
- **Label noise in adversarial training comes from the mismatch between the true label distribution and assigned label distribution**

$$P(Y' \neq \tilde{Y}'|x') \geq \|P(Y'|x') - P(\tilde{Y}'|x')\|_{\mathrm{TV}}$$

- **Instead of one-hot labels, we can use the predictive label distribution of an adversarially robust classifier to label adversarial examples, which can approximate the true label distribution of adversarial examples**

Mitigate label noise in adversarial training

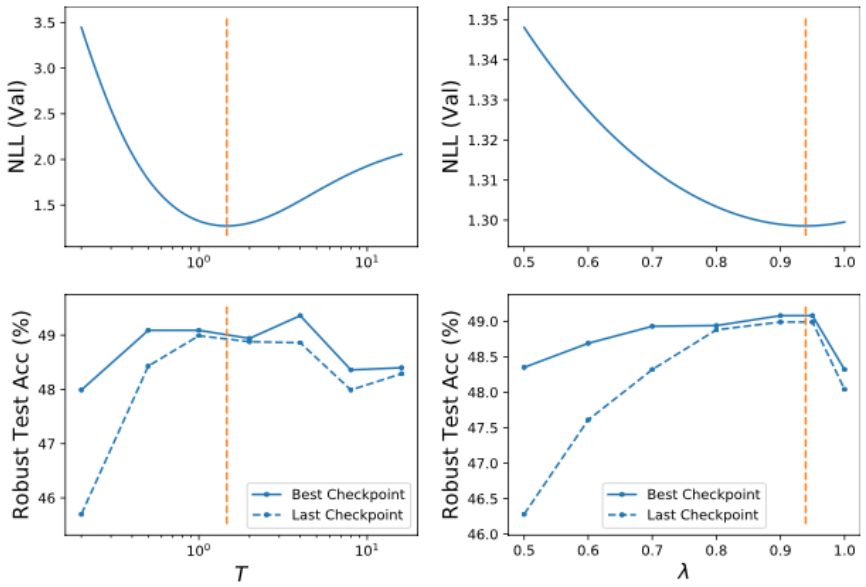
- **We can use <u>confidence calibration</u> to further improve the approximation**
  - **temperature scaling**
  - **interpolation between predictive distribution and one-hot label**



Approximation error of the true label distribution

Mitigate label noise in adversarial training

- **We can use confidence calibration to further improve the approximation**
  - **temperature scaling**
  - **interpolation between predictive distribution and one-hot label**



Approximation error of the true label distribution

Overfitting gap: difference between the robust accuracies at the best and last checkpoint

Future works

- Other effects of label noise on adversarial training
- Advanced methods to mitigate label noise in adversarial training
- …