# On the Impact of Hard Adversarial Instances on Overfitting in Adversarial Training

Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, Sabine Süsstrunk

Journal Track on ICML 2025

# Outline

- We study robust overfitting issue in adversarial training.

Theoretical contribution by analyses:

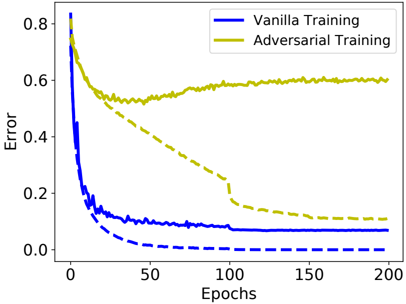
- In general cases, hard adversarial instances lead to more severe overfitting.

Empirical contribution by case studies:

- Downplaying hard adversarial instances help mitigate adversarial overfitting.

# Adversarial Overfitting

- We use the average training loss per instance to define the difficulty of a training instance, and then monitor how loss evolves during training.



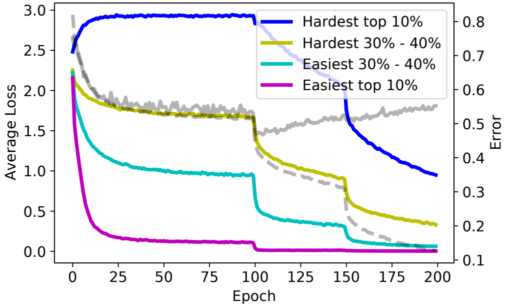




Figure: (Left) Learning curves of vanilla training and adversarial training. (Right) Learning curves of training instances of different difficulty levels. The grey curves are overall learning curves for reference.

# Theoretical Analyses: Why?

**Data** The data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is binary, i.e., $\boldsymbol{x}_i \in \mathbb{R}^m, y_i \in \{-1, +1\}$. It is sub-Gaussian with positive conditional variance $\sigma^2 = \mathbb{E}[Var[y|\boldsymbol{x}]] = \sigma^2 > 0$.

# Theoretical Analyses: Why?

**Data** The data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is binary, i.e., $\boldsymbol{x}_i \in \mathbb{R}^m, y_i \in \{-1, +1\}$. It is sub-Gaussian with positive conditional variance $\sigma^2 = \mathbb{E}[Var[y|\boldsymbol{x}]] = \sigma^2 > 0$.

Lipschitz constant $Lip(f(\cdot, \theta)) = sup_{\boldsymbol{x}_1, \boldsymbol{x}_2} \frac{\|f(\boldsymbol{x}_1, \theta) - f(\boldsymbol{x}_2, \theta)\|}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|}$ is a good indicator of the adversarial vulnerability.

# Theoretical Analyses: Why?

## Theorem (Informal and Simplified)

*Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, and a model parameterized by bounded parameters $\theta$, we conduct adversarial training and let $\boldsymbol{x}'$ to the adversarial examples of $\boldsymbol{x}$. If the training loss $C = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}'_i, \theta) - y_i)^2$ is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

$\sigma\uparrow$, $H\uparrow$; $\epsilon\uparrow$, $H\uparrow$; $C\downarrow$, $H\uparrow$.

# Theoretical Analyses: Why?

## Theorem (Informal and Simplified)

*Given training data* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, *and a model parameterized by bounded parameters* $\theta$, *we conduct adversarial training and let* $\boldsymbol{x}'$ *to the adversarial examples of* $\boldsymbol{x}$. *If the training loss* $C = \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}'_i, \theta) - y_i)^2$ *is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

- Lipschitz constant indicates adversarial vulnerability. [1]

[1] L. Weng, et. al. "Towards fast computation of certified robustness for relu networks".

# Theoretical Analyses: Why?

### Theorem (Informal and Simplified)

*Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, and a model parameterized by bounded parameters $\theta$, we conduct adversarial training and let $\boldsymbol{x}'$ to the adversarial examples of $\boldsymbol{x}$. If the training loss $C = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}'_i, \theta) - y_i)^2$ is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

- Lipschitz constant indicates adversarial vulnerability. [1]
  - $C$ is sufficiently small $\Longrightarrow$ Lipschitz constant indicates generalization gap.

[1] L. Weng, et. al. "Towards fast computation of certified robustness for relu networks".

# Theoretical Analyses: Why?

## Theorem (Informal and Simplified)

*Given training data* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$*, and a model parameterized by bounded parameters* $\theta$*, we conduct adversarial training and let* $\boldsymbol{x}'$ *to the adversarial examples of* $\boldsymbol{x}$*. If the training loss* $C = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}'_i, \theta) - y_i)^2$ *is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

- Lipschitz constant indicates adversarial vulnerability. [1]
  - $C$ is sufficiently small $\Longrightarrow$ Lipschitz constant indicates generalization gap.
  - $C\downarrow$: training processes $\Longrightarrow$ $H\uparrow$: overfitting.

[1] L. Weng, et. al. "Towards fast computation of certified robustness for relu networks".

# Theoretical Analyses: Why?

## Theorem (Informal and Simplified)

*Given training data* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, *and a model parameterized by bounded parameters* $\theta$, *we conduct adversarial training and let* $\boldsymbol{x}'$ *to the adversarial examples of* $\boldsymbol{x}$. *If the training loss* $C = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i', \theta) - y_i)^2$ *is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

- Lipschitz constant indicates adversarial vulnerability. [1]
  - $C$ is sufficiently small $\Longrightarrow$ Lipschitz constant indicates generalization gap.
  - $C \downarrow$: training processes $\Longrightarrow$ $H \uparrow$: overfitting.
  - $\sigma \uparrow$: harder instances $\Longrightarrow$ $H \uparrow$: overfitting.

[1] L. Weng, et. al. "Towards fast computation of certified robustness for relu networks".

# Theoretical Analyses: Why?

## Theorem (Informal and Simplified)

*Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and a model parameterized by bounded parameters $\theta$, we conduct adversarial training and let $\boldsymbol{x}'$ to the adversarial examples of $\boldsymbol{x}$. If the training loss $C = \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}'_i, \theta) - y_i)^2$ is sufficiently small, then the Lipschitz constant of the model is lower bounded by the following equation almost surely.*

$$Lip(f(\cdot, \theta)) \geq H(\sigma^2, \epsilon, C)$$

- Lipschitz constant indicates adversarial vulnerability. [1]
  - $C$ is sufficiently small $\Longrightarrow$ Lipschitz constant indicates generalization gap.
  - $C \downarrow$: training processes $\Longrightarrow H \uparrow$: overfitting.
  - $\sigma \uparrow$: harder instances $\Longrightarrow H \uparrow$: overfitting.
  - $\epsilon \uparrow$: larger adversarial budget $\Longrightarrow H \uparrow$: overfitting.

[1] L. Weng, et. al. "Towards fast computation of certified robustness for relu networks".

# Empirical Observation: How?

- Methods mitigating adversarial overfitting implicitly downplay hard instances.
  - Weaker perturbation; adaptive and easier targets; smaller weights.
- Methods highlighting hard instances do not achieve true robustness.

# Empirical Observation: How?

- Methods mitigating adversarial overfitting implicitly downplay hard instances.
  - Weaker perturbation; adaptive and easier targets; smaller weights.
- Methods highlighting hard instances do not achieve true robustness.

**Contributions:**

**Theory-backed analysis of adversarial overfitting in the lens of training data.**

Thanks!