



ICML
International Conference
On Machine Learning



SORA: Free Second-Order Attacks in Fast Adversarial Training

International Conference on Machine Learning 2026

Mazdak Teymourian*, Ramtin Moslemi*,
Farzan Rahmani, Mohammad Hossein Rohban

Robust and Interpretable Machine Learning Lab @ Sharif University of Technology

Seoul, South Korea | July 2026

Adversarial Training

- Mądry et al.¹ formulate the min–max optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \mathcal{L}(f_{\theta}(x + \delta), y) \right]$$

- Mądry et al.¹ use the **expensive** multi-step attack PGD.
- We can use the **Fast** Gradient Sign Method (FGSM)² instead:

$$\delta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f_{\theta}(x), y)),$$

but there is a problem ...

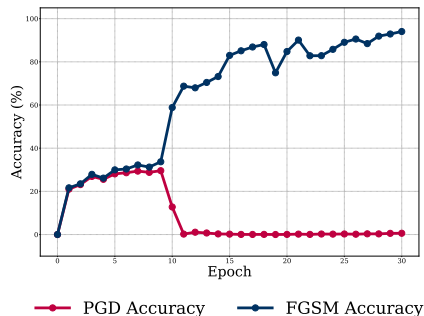
¹Mądry et al. [Towards Deep Learning Models Resistant to Adversarial Attacks](#). **ICLR**, 2018.

²Goodfellow et al. [Explaining and Harnessing Adversarial Examples](#). **ICLR**, 2015.

Catastrophic Overfitting

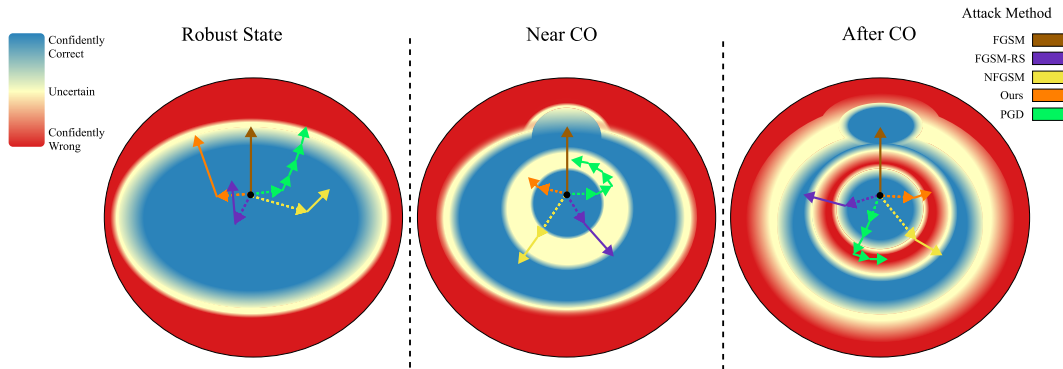
Wong et al.³ identified a critical failure mode in single-step AT.

- Single-step (FGSM) accuracy rises
- Multi-step (PGD) accuracy drops
- This happens very quickly
- Wong et al.³ call this **Catastrophic Overfitting**

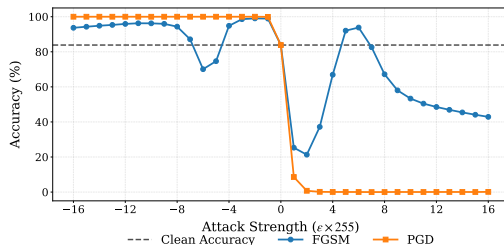


³Wong et al. **Fast is Better than Free: Revisiting Adversarial Training**. **ICLR**, 2020.

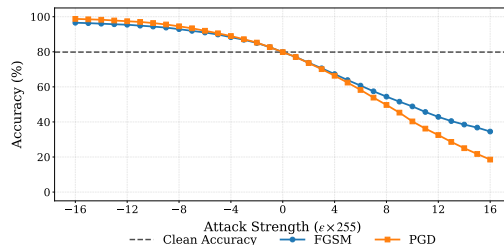
Evolution of the Loss Landscape Geometry



Epsilon Overfitting



(a) FGSM Adversarial Training



(b) SORA Adversarial Training

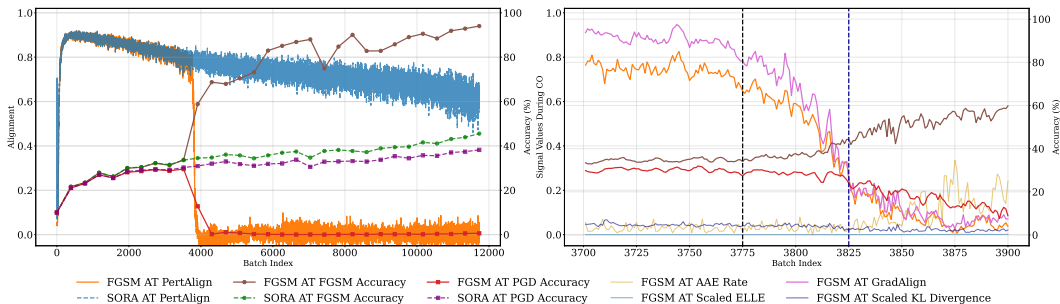
Perturbation Alignment

- Formally, **PertAlign** is defined as:

$$\cos(\nabla_x \mathcal{L}(f_\theta(\mathbf{x}), y), \nabla_x \mathcal{L}(f_\theta(\mathbf{x} + \delta), y))$$

- This measures the alignment between the loss gradient with respect to:
 - The original sample
 - The adversarial example

PertAlign in Action



Our General Idea

- If we have **high PertAlign**:
 - Loss surface is locally linear
 - Multi- and single-step methods yield similar δ
 - We can increase the step-size
- If we have **low PertAlign**:
 - Loss surface is distorted
 - Multi- and single-step methods yield very different δ
 - We must reduce the step-size

Second-Order Adaptive Method (SORA)

- SORA has two main components:

- 1 Adaptive maximum step-size:

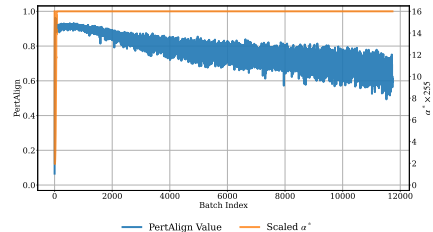
$$\alpha^* = \begin{cases} \min \left(\alpha_{\max}, \frac{\alpha_0}{1 - \frac{p^T g'}{\|g\|_1}} \right), & \frac{p^T g'}{\|g\|_1} < 1, \\ \alpha_{\max}, & \text{otherwise.} \end{cases}$$

- 2 Per-pixel channel diversification:

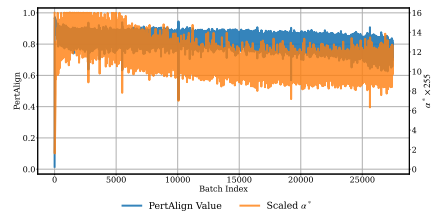
$$\alpha \sim \mathcal{U}(0, \alpha^*)^d$$

- We can then generate adversarial examples:

$$x' \leftarrow x + \eta + \alpha \odot \text{sign}(g)$$

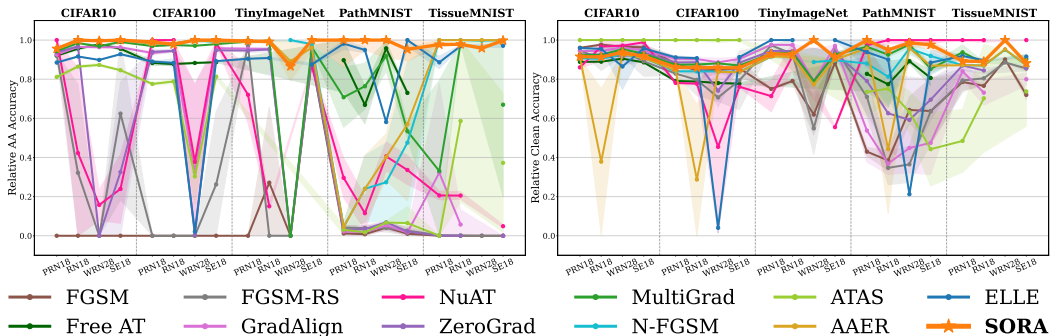


(a) CIFAR-10

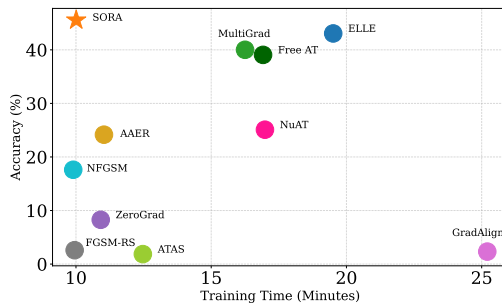


(b) IMAGENET-100

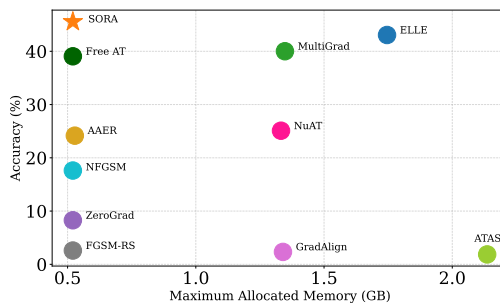
Performance and Generalization



Training Costs



(a) Time



(b) Memory

Figure: Training time and memory usage on PATHMNIST with PreActResNet-18, trained for 30 epochs, measured on an NVIDIA GeForce RTX 4090 GPU. The ★ marks SORA. The vertical axis in both figures represents PGD-10 accuracy.

Contributions

- Understanding CO through the new lens of **Epsilon Overfitting**.
- Introducing **PertAlign** as a CO tracking metric with no overhead.
- Formalizing the characteristics of a **good solution**:
 - Robustness across datasets and architectures
 - High clean and robust accuracy with low cost
 - Dataset and architecture agnostic hyperparameters
- Introducing **SORA** as a reliable single-step AT method.

Read Our Paper

- Theoretical insights
- Reinterpreting previous work
- Ablations
- Euclidean norm attacks
- Vision Transformers



Thank you for listening!

Presenters: {[mazdak.teymourian01](#), [ramtin.moslemi](#)}@sharif.edu

Corresponding Author: [rohban@sharif.edu](#)

