Friendly Adversarial Training:
Attacks Which Do Not Kill Training Make Adversarial Learning Stronger

Jingfeng Zhang[1*], Xilie Xu[2*], Bo Han[34], Gang Niu[4], Lichen Cui[5], Masashi Sugiyama[46], and Mohan Kankanhalli[1]

[1]Department of Computer Science, National University of Singapore

[2]Taishan Colleague, Shandong University

[3]Department of Computer Science, Hong Kong Baptist University

[4]RIKEN Center for Advanced Intelligence Project

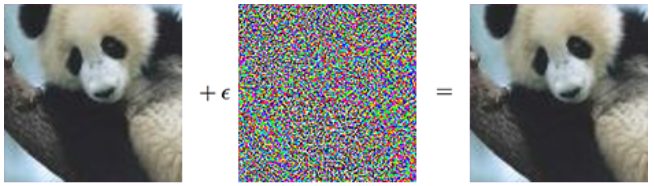
[5]School of Software & C-FAIR, Shandong University

[6]Graduate School of Frontier Sciences, The University of Tokyo

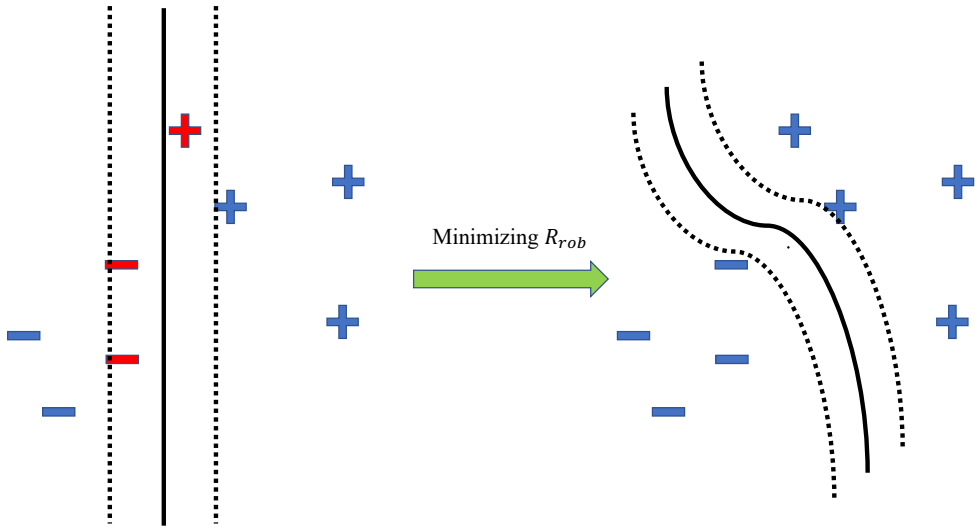
Virtual ICML 2020 July, 2020

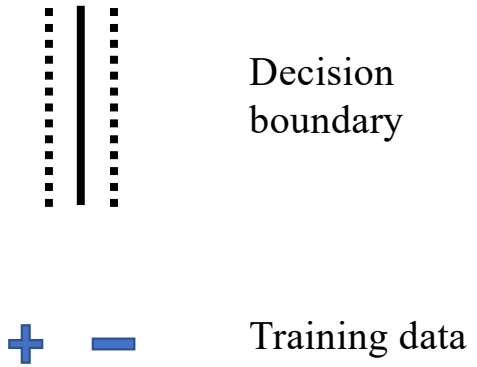
Purpose of adversarial learning

https://blog.openai.com/adversarial-example-research/



- **Adversarial data** can easily fool the standard trained classifier.
- **Adversarial training** so far is the most effective method for obtaining the adversarial robustness of the trained classifier.





Purpose 1: correctly classify the data.
Purpose 2: make the decision boundary thick so that no data is encouraged to fall inside the decision boundary.

Conventional formulation of adversarial training

- Minimax formulation:

$$\min_{f \in \mathcal{F}} \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \ell(f\ (\widetilde{x_i})\ , y_i), \text{ where } \widetilde{x_i} = argmax_{x \in B(x_i)}\ \ell(f(\tilde{x}), y_i)$$

Outer minimization Inner maximization

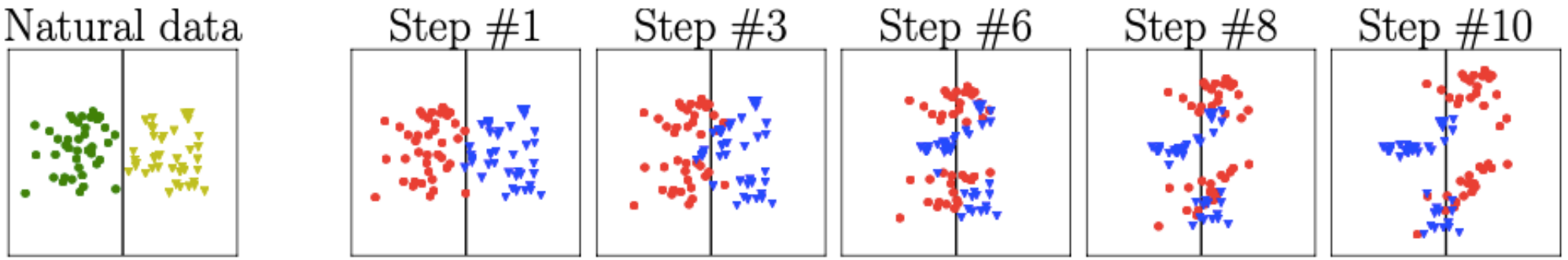
- Projected gradient descent (PGD) – adversarial training approximately realizes this minimax formulation.
- PGD formulates the problem of finding **the most adversarial data** as a constrained optimization problem. Namely, given a starting point $x^{(0)} \in \mathcal{X}$ and step size α, PGD works as followed:

$$x^{(t+1)} = \Pi_{B\left(x^{(0)}\right)} \left(x^{(t)} + \alpha\ sign\left(\nabla_{x^{(t)}}\ \ell\left(f_\theta\left(x^{(t)} \right), y \right) \right) \right), \mathrm{t} \in N$$

The minimax formulation is pessimistic.

- Many existing studies found the minimax-based adversarial training causes the severe degradation of the natural generalization. Why?

The adversarial data generated by PGD



The cross-over mixture problem!

Is the minimax formulation suitable to the adversarial training?

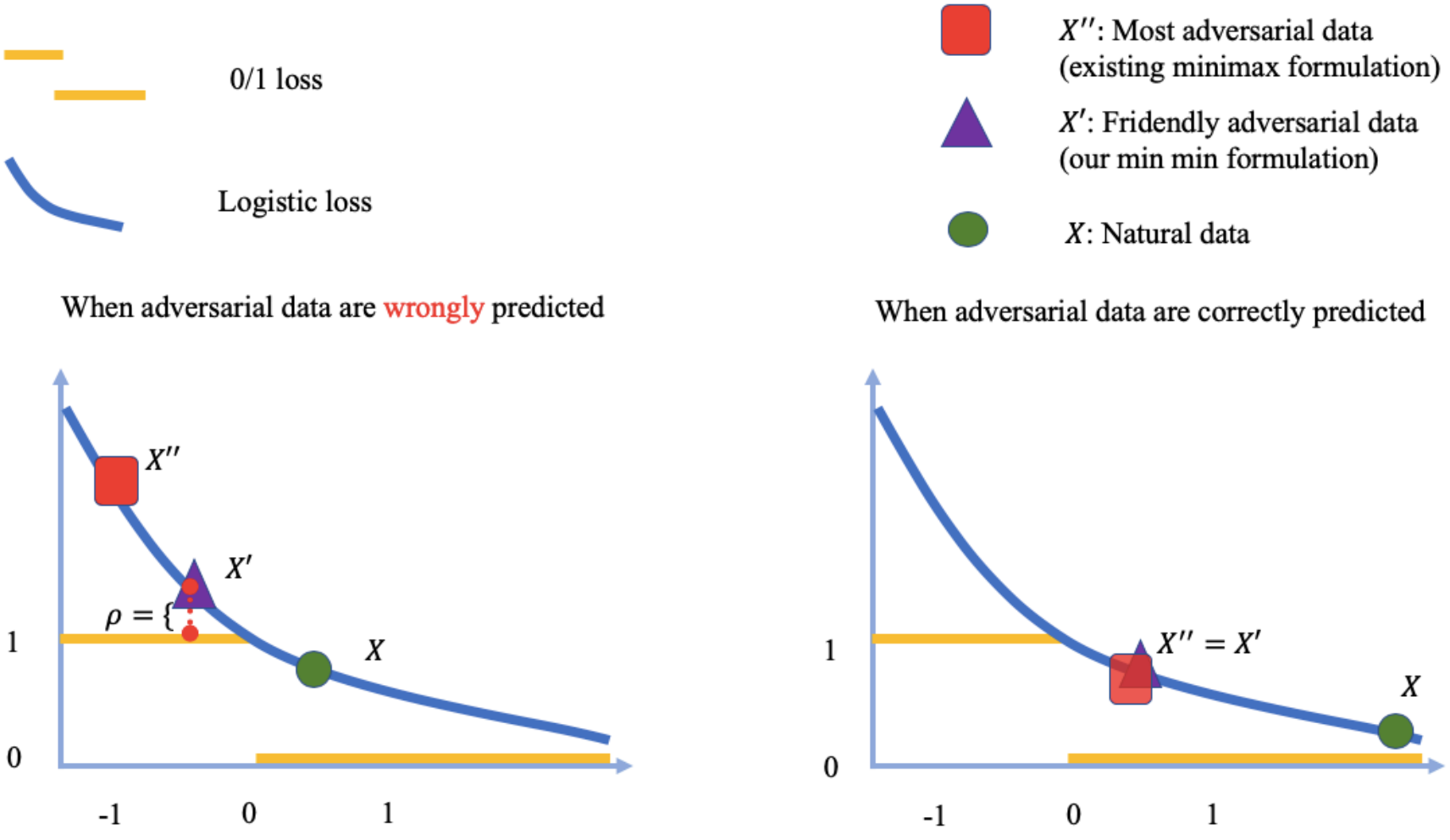
Min-min formulation for the adversarial training

- The outer minimization keeps the same. Instead of generating adversarial data $\widetilde{x_i}$ via inner maximization, we generate $\widetilde{x_i}$ as follows:

$$\widetilde{x_i} = \arg min_{\tilde{x} \in B(x_i)} \ell(f(\tilde{x}), y_i) \text{ s.t. } \ell(f(\tilde{x}), y_i) - \min_{y \in \mathcal{Y}} \ell(f(\tilde{x}), y_i) \geq \rho$$

- The constraint firstly ensures $y_i \neq \arg\min_{y \in \mathcal{Y}} \ell(f(\tilde{x}), y_i)$ or $\tilde{x}$ is misclassified, and secondly ensures the wrong prediction of $\tilde{x}$ is better than the desired prediction y_i by at least the margin ρ in terms of the loss value.

Adversarial data by min-min and minimax formulation



A tight upper bound on the adversarial risk

The adversarial risk $\mathfrak{R}_{rob}(f) \coloneqq \mathbb{E}_{(X,Y\in D)}\mathbb{1}\{\exists\, X' \in B(X): f(X') \neq Y\}$

Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." ICML 2019

Minimizing the adversarial risk captures the two purposes of the adversarial training:
(a) correctly classify the natural data and (b) make the decision boundary thick.

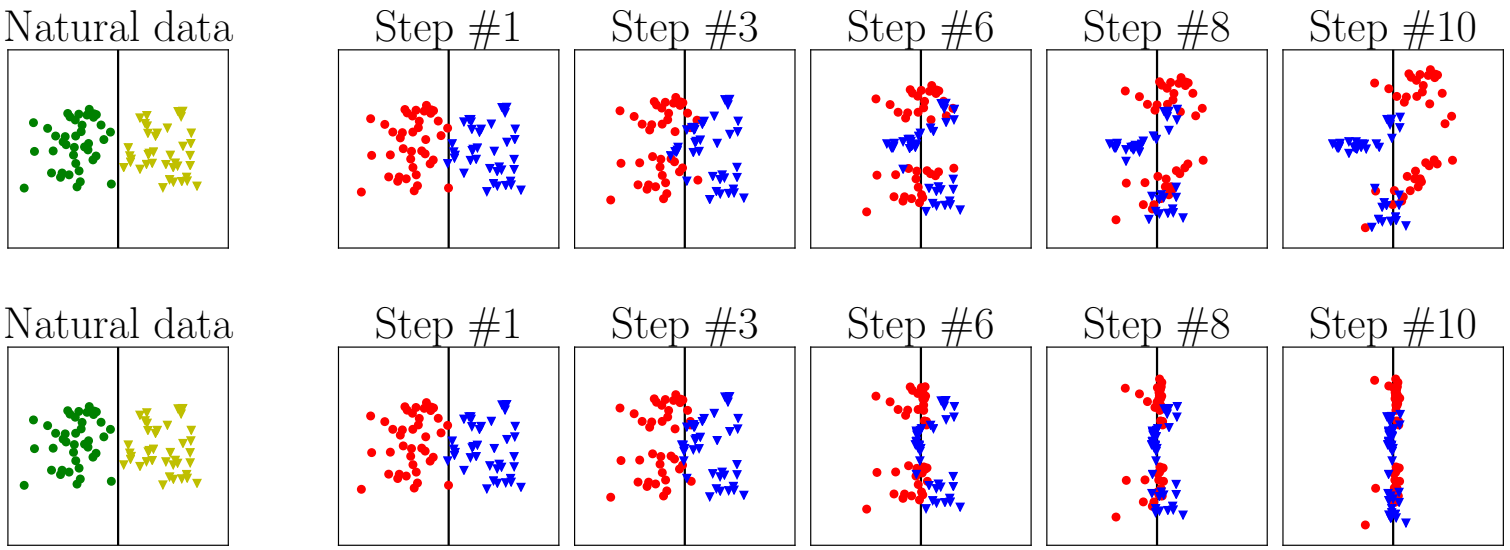
Theorem 1. For any classifier f, any non-negative surrogate loss function ℓ which upper bounds the 0/1 loss, and any probability distribution $\mathcal{D}$, we have

$$\mathcal{R}_{\text{rob}}(f) \leq \underbrace{\mathbb{E}_{(X,Y)\sim\mathcal{D}}\ell(f(X),Y)}_{\text{For standard test accuracy}} + \underbrace{\mathbb{E}_{(X,Y)\sim\mathcal{D},X'\in\mathcal{B}_\epsilon[X,\epsilon]}\ell^*(f(X'),Y)}_{\text{For robust test accuracy}},$$

where

$$\ell^* = \begin{cases} \min \ell(f(X'),Y) + \rho, & \text{if } f(X') \neq Y, \\ \max \ell(f(X'),Y), & \text{if } f(X') = Y. \end{cases}$$

Realization of our min-min formulation – friendly adversarial training (FAT)



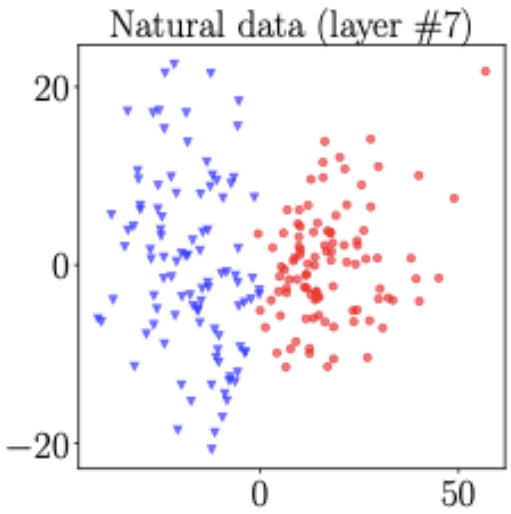
Conventional PGD generating most adversarial data

Early stopped PGD (ours) generating friendly adversarial data

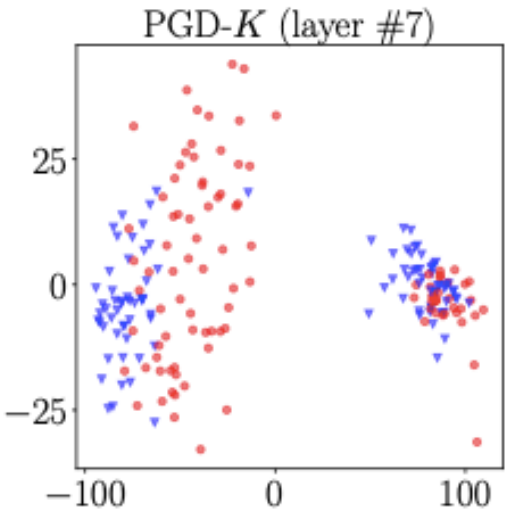
Friendly adversarial training (FAT) employs the **friendly adversarial data** generated by **early stopped PGD** to update the model.

Benefits (a): Alleviate the cross-over mixture problem

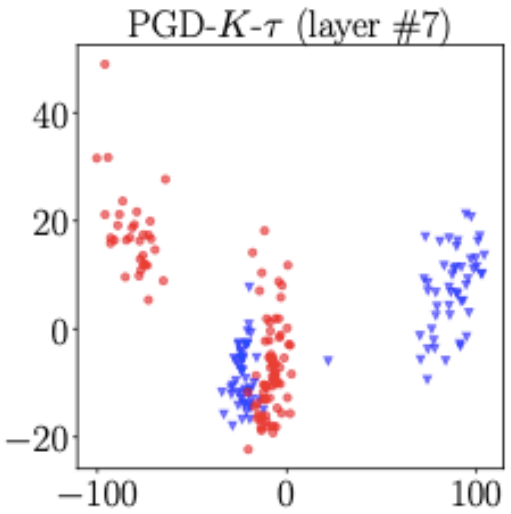
- In the classification of the CIFAR-10 dataset, the cross-over mixture problem may not appear in the input space, but in the middle layers.



Natural data
(not mixed)

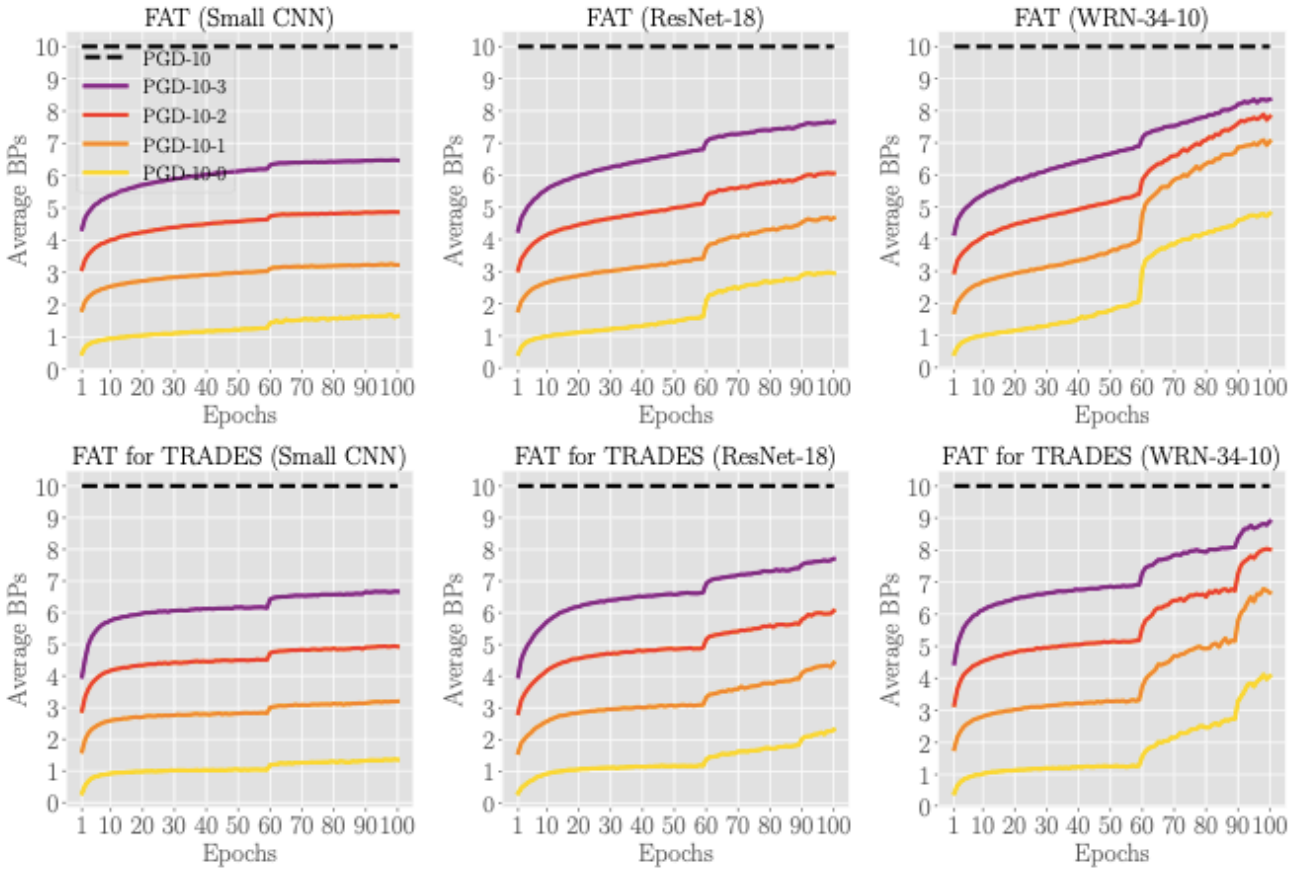


Most adversarial data
generated by
conventional PGD
(significantly mixed)



Friendly adversarial data
generated by
early stopped PGD (not
significantly mixed)

Benefits (b): FAT is computationally efficient.

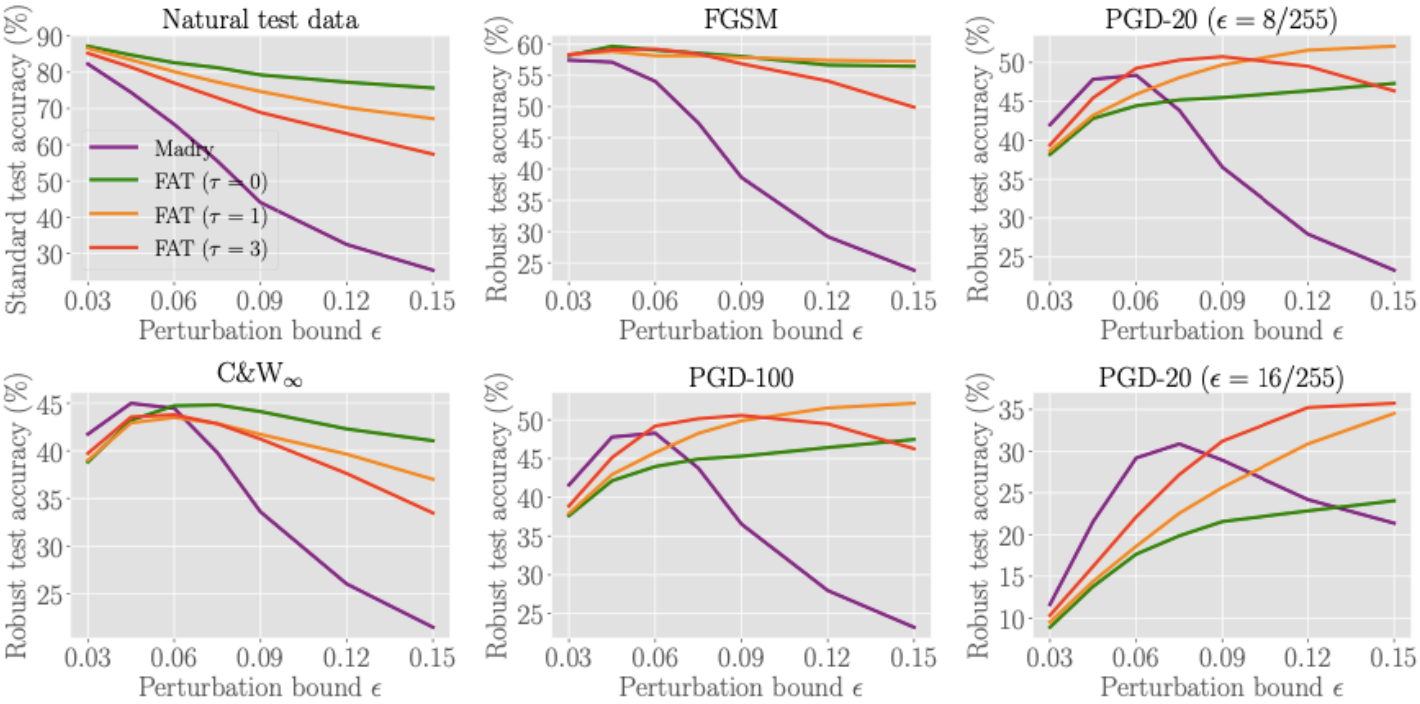


We report the average backward propagations (BPs) per epoch over training process.

Dashed line is existing adversarial training based on conventional PGD.

Solid lines are friendly adversarial trainings based on early stopped PGD.

Benefits (c): FAT can enable larger defense parameter ϵ_{train}



For CIFAR-10 dataset, we adversarially train deep neural networks with $\epsilon_{train} \in [0.03, 0.15]$, and evaluate each robust model with 6 evaluation metrics (1 natural generalization metric + 5 robustness metrics)

The purple line represents existing adversarial training.

The red, orange and green lines represent our friendly adversarial training with different configurations.

Benefits (d): Benchmarking on Wide ResNet.

Table 1: Evaluations (test accuracy) of deep models (WRN-34-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$
Madry	87.30	56.10	45.80	46.80
CAT	77.43	57.17	46.06	42.28
DAT	85.03	63.53	48.70	47.27
FAT ($\epsilon = 8/255$)	**89.61** $\pm$ 0.329	65.19 $\pm$ 0.269	46.45 $\pm$ 0.448	46.81 $\pm$ 0.308
FAT ($\epsilon = 16/255$)	87.02 $\pm$ 0.212	**65.72** $\pm$ 0.296	**49.77** $\pm$ 0.177	**48.59** $\pm$ 0.314

Results of Madry, CAT and DAT are reported in [14]. FAT has the same evaluations.

Table 2: Evaluations (test accuracy) of deep models (WRN-34-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$
TRADES ($\beta = 1.0$)	88.64	56.38	49.14	-
FAT for TRADES ($\epsilon = 8/255$)	**89.94** $\pm$ 0.303	**61.00** $\pm$ 0.418	**49.70** $\pm$ 0.653	49.35 $\pm$ 0.363
TRADES ($\beta = 6.0$)	84.92	61.06	56.61	**54.47**
FAT for TRADES ($\epsilon = 8/255$)	**86.60** $\pm$ 0.548	**61.97** $\pm$ 0.570	55.98 $\pm$ 0.209	54.29 $\pm$ 0.173
FAT for TRADES ($\epsilon = 16/255$)	84.39 $\pm$ 0.030	61.73 $\pm$ 0.131	**57.12** $\pm$ 0.233	54.36 $\pm$ 0.177

Results of TRADES ($\beta = 1.0$ and 6.0) are reported in [13]. FAT for TRADES has the same evaluations.

FAT can improve standard test accuracy while maintain the superior adversarial robustness.

[14] Wang, Yisen, et al. "On the convergence and robustness of adversarial training." ICML 2019

[13] Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." ICML 2019

Conclusion and future work

- We propose a novel min-min formulation for adversarial training.
- Friendly adversarial training (FAT) to realize this min-min formulation.
- FAT helps alleviate the problem of cross-over mixture.
- FAT is computationally efficient.
- FAT can enable larger perturbation bounds ϵ_{train}.
- FAT can achieve competitive performance on the large capacity networks.

- Besides FAT, one of the potential future work is to find a better realization of our min-min formulation.

Thanks for your interest in our work.