YOU ONLY PROPAGATE ONCE

YIPING LU PEKING UNIVERSITY

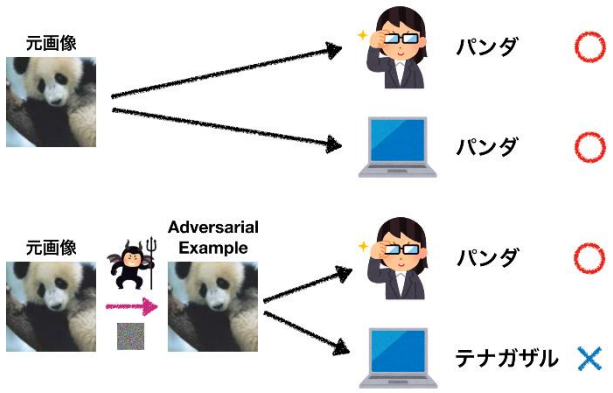
TOWARDS DEEP LEARNING MODELS RESISTANT TO ADVERSARIAL ATTACKS

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

Robust Optimization

Problem:

- More capacity and stronger adversaries decrease transferability. Always 10 times wider
- PGD training is expansive!



Can adversarial training cheaper

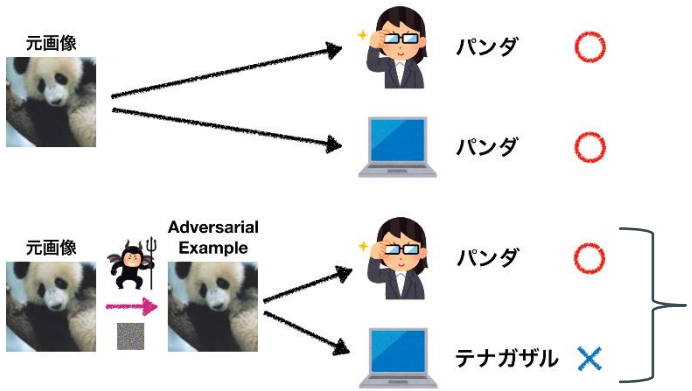
TOWARDS DEEP LEARNING MODELS RESISTANT TO ADVERSARIAL ATTACKS

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

Robust Optimization

Problem:

- More capacity and stronger adversaries decrease transferability. Always 10 times wider
- PGD training is expansive!

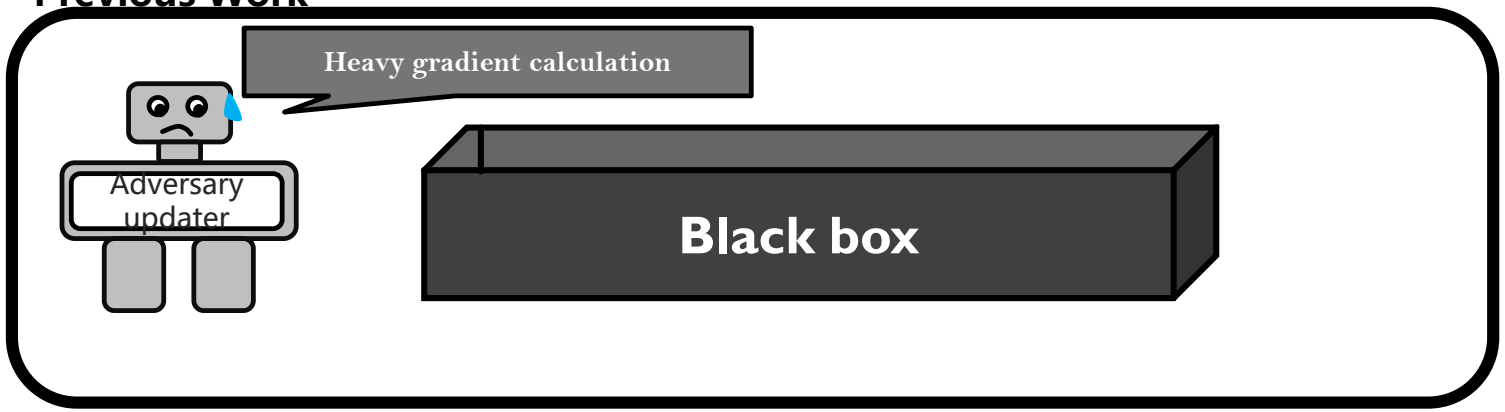


Solving the maximization needs BP many times

FGSM adversarial example is too easy.

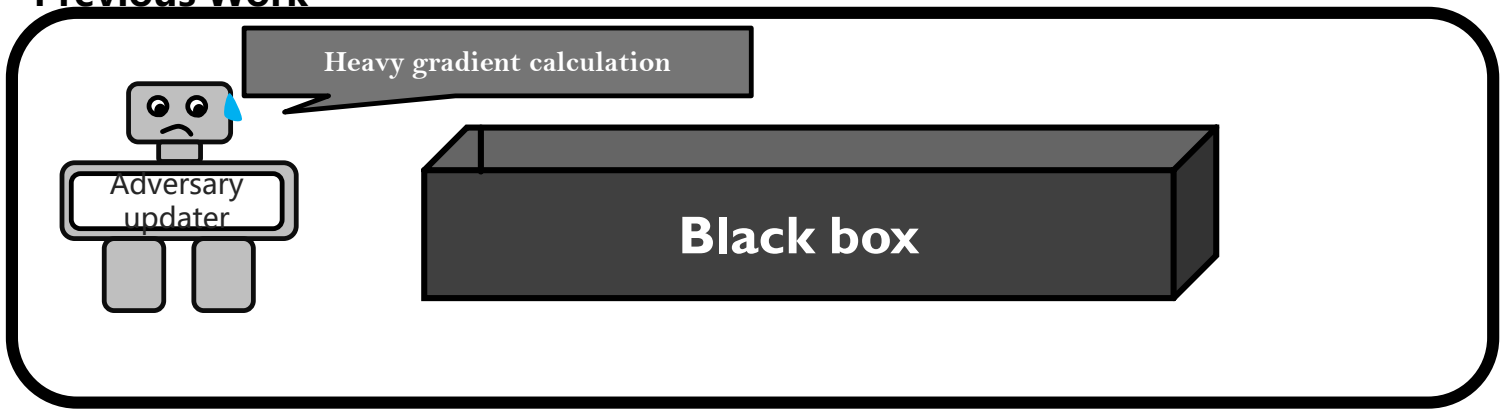
TAKE NEURAL NETWORK ARCHITECTURE INTO CONSIDERATION

Previous Work

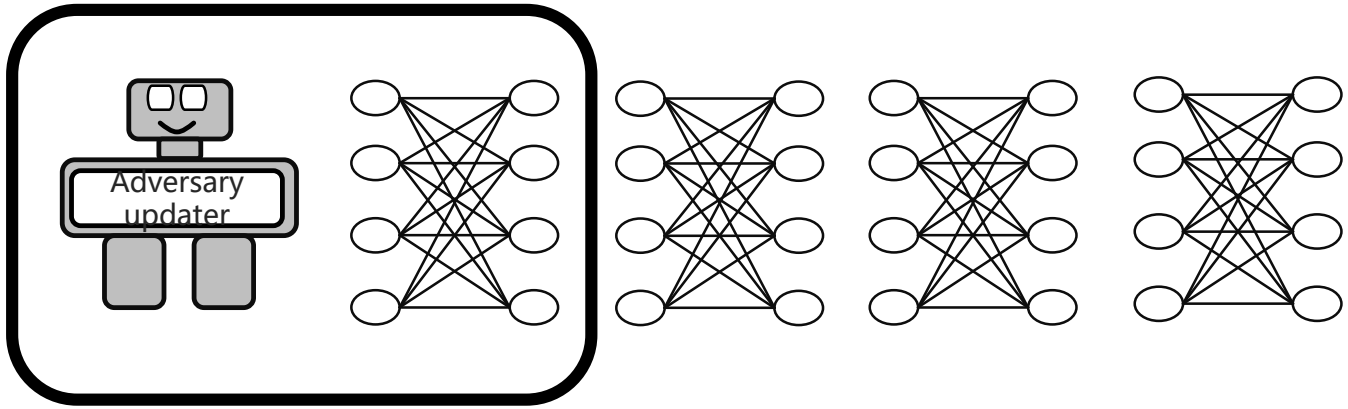


TAKE NEURAL NETWORK ARCHITECTURE INTO CONSIDERATION

Previous Work



YOPO



Take **The Neural Network Structure** Into Consideration

DIFFERENTIAL GAME

$$\min_{\theta} \max_{\|\eta\|_\infty \leq \epsilon} J(\theta, \eta) := \frac{1}{N} \sum_{i=1}^{N} \ell_i(x_{i,T}) + \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} R_t(x_{i,t}; \theta_t) \tag{2}$$

$$\text{subject to } x_{i,1} = f_0(x_{i,0} + \eta; \theta_0), i = 1, 2, \cdots, N$$

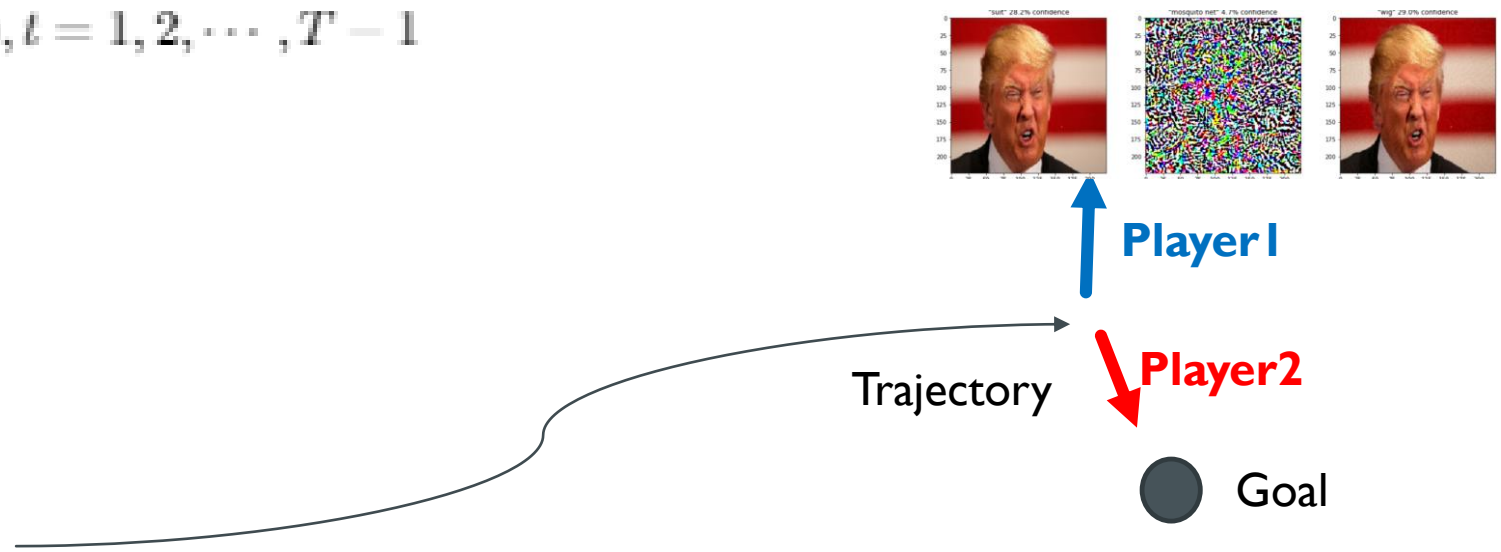
$$x_{i,t+1} = f_t(x_{i,t}, \theta_t), t = 1, 2, \cdots, T-1$$

DIFFERENTIAL GAME

$$\min_{\theta} \max_{\|\eta\|_\infty \leq \epsilon} J(\theta, \eta) := \frac{1}{N} \sum_{i=1}^{N} \ell_i(x_{i,T}) + \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} R_t(x_{i,t}; \theta_t) \quad (2)$$

$$\text{subject to } x_{i,1} = f_0(x_{i,0} + \eta; \theta_0), i = 1, 2, \cdots, N$$

$$x_{i,t+1} = f_t(x_{i,t}, \theta_t), t = 1, 2, \cdots, T-1$$



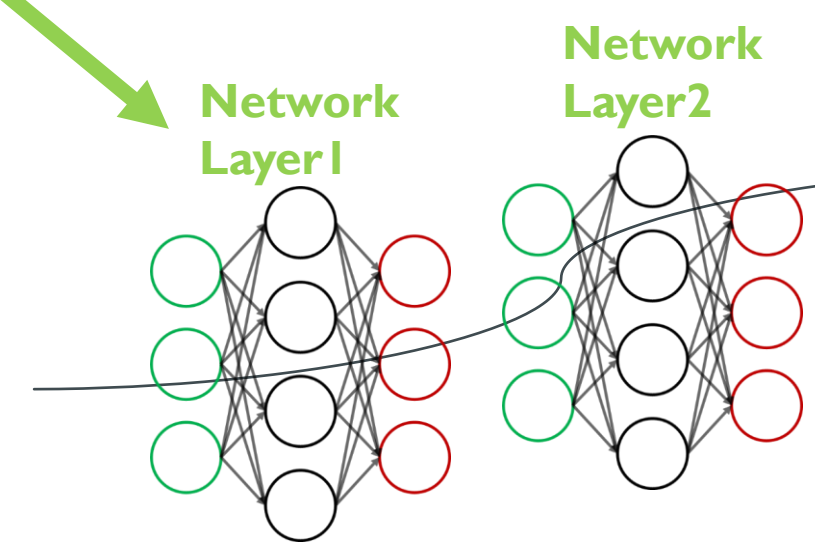
DIFFERENTIAL GAME

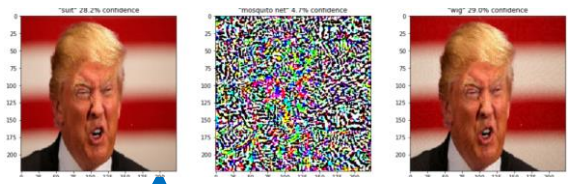
$$\min_{\theta} \max_{\|\eta\|_\infty \le \epsilon} J(\theta, \eta) := \frac{1}{N} \sum_{i=1}^{N} \ell_i(x_{i,T}) + \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} R_t(x_{i,t}; \theta_t)$$

$$\text{subject to } x_{i,1} = f_0(x_{i,0} + \eta; \theta_0), i = 1, 2, \cdots, N$$
$$x_{i,t+1} = f_t(x_{i,t}, \theta_t), t = 1, 2, \cdots, T-1$$

(2)

Composition Structure





Player1

Player2

Trajectory

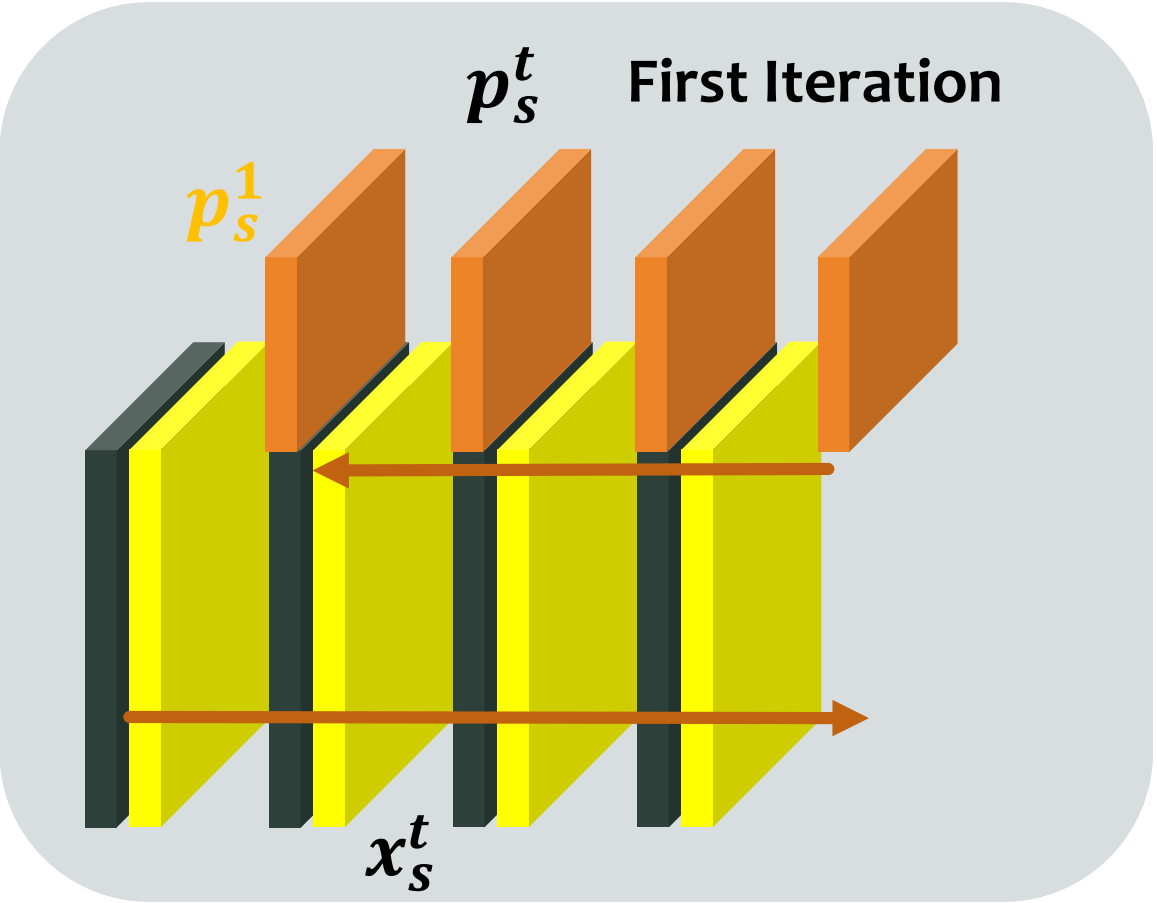
Goal

DECOUPLE TRAINING

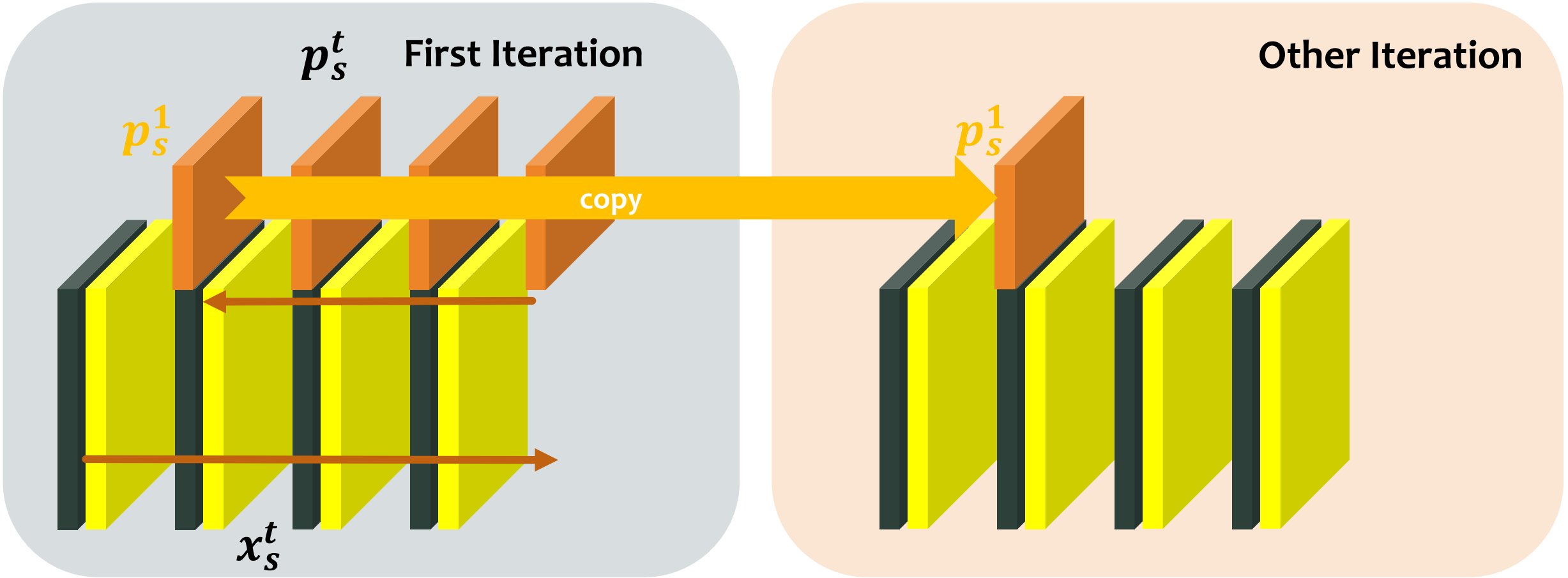
- Synthetic gradients [Jaderberg et al.2017]
- Lifted Neural Network [Askari et al.2018] [Gu et al.2018] [li et al.2019]
- Delayed Gradient [Huo et al.2018]
- Block Coordinate Descent Approach [Lau et al. 2018]

- Can Control perspective helps us to understand decoupling?
- Our idea: Decouple the gradient back propagation with the adversary updating.

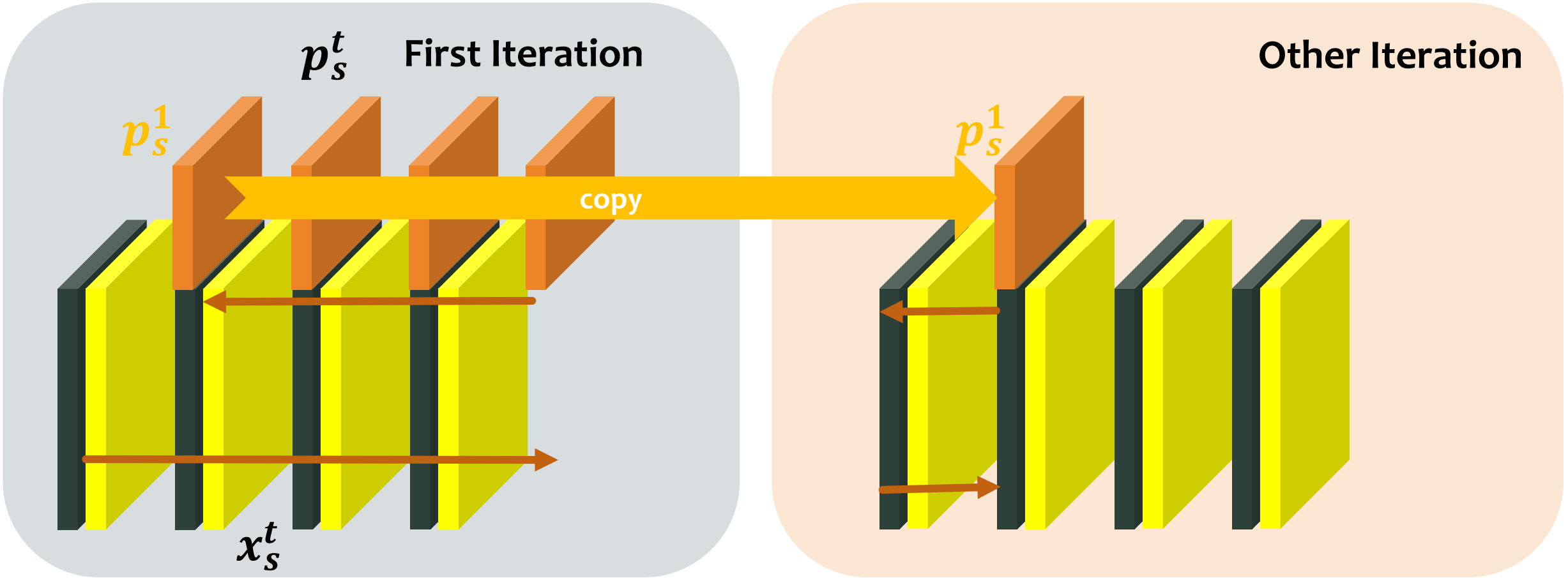
YOPO(YOU ONLY PROPAGATE ONCE)



YOPO(YOU ONLY PROPAGATE ONCE)



YOPO(YOU ONLY PROPAGATE ONCE)



WHY DECOUPLING

Theorem 1. *(PMP for adversarial defense) There exists co-state processes* $p_s^* := p_{s,t}^* : t = 0, \cdots, T$ *such that the following holds for all* $t \in [T]$ *and* $s \in [S]$*:*

$$x_{s,t+1}^* = \nabla_p H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad x_{s,0}^* = x_{s,0} + \eta \tag{5}$$

$$p_{s,t}^* = \nabla_x H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad p_{s,T}^* = -\frac{1}{S}\nabla\Phi(x_{s,T}^*) \tag{6}$$

At the same the the parameter of the first layer θ_0^* *satisfies*

$$\sum_{s=1}^{S} H_t(x_{s,0} + \hat{\eta}, p_{s,t+1}^*, \theta_0^*), \forall\theta \in \Theta_t \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta_0^*) \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta), \forall\theta \in \Theta_0, \|\hat{\eta}\|_\infty \leq \epsilon \tag{7}$$

and parameter of the other layers $\theta_t^*, t = 1, 2, \cdots, T$ *will maximize the Hamiltonian functions*

$$\sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*) \geq \sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta), \forall\theta \in \Theta_t \tag{8}$$

WHY DECOUPLING

Theorem 1. *(PMP for adversarial defense) There exists co-state processes* $p_s^* := p_{s,t}^* : t = 0, \cdots, T$ *such that the following holds for all* $t \in [T]$ *and* $s \in [S]$*:*

$$x_{s,t+1}^* = \nabla_p H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad x_{s,0}^* = x_{s,0} + \eta \tag{5}$$

Forward Propagation

$$p_{s,t}^* = \nabla_x H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad p_{s,T}^* = -\frac{1}{S}\nabla\Phi(x_{s,T}^*) \tag{6}$$

At the same the the parameter of the first layer θ_0^* *satisfies*

$$\sum_{s=1}^{S} H_t(x_{s,0} + \hat{\eta}, p_{s,t+1}^*, \theta_0^*), \forall\theta \in \Theta_t \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta_0^*) \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta), \forall\theta \in \Theta_0, \|\hat{\eta}\|_\infty \leq \epsilon \tag{7}$$

and parameter of the other layers $\theta_t^*, t = 1, 2, \cdots, T$ *will maximize the Hamiltonian functions*

$$\sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*) \geq \sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta), \forall\theta \in \Theta_t \tag{8}$$

WHY DECOUPLING

Theorem 1. *(PMP for adversarial defense) There exists co-state processes* $p_s^* := p_{s,t}^* : t = 0, \cdots, T$ *such that the following holds for all* $t \in [T]$ *and* $s \in [S]$*:*

$$x_{s,t+1}^* = \nabla_p H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad x_{s,0}^* = x_{s,0} + \eta \tag{5}$$

Gradient of feature map

$$p_{s,t}^* = \nabla_x H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad p_{s,T}^* = -\frac{1}{S}\nabla\Phi(x_{s,T}^*) \tag{6}$$

Backward Propagation

At the same the the parameter of the first layer θ_0^* *satisfies*

$$\sum_{s=1}^{S} H_t(x_{s,0} + \hat{\eta}, p_{s,t+1}^*, \theta_0^*), \forall\theta \in \Theta_t \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta_0^*) \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta), \forall\theta \in \Theta_0, \|\hat{\eta}\|_\infty \leq \epsilon \tag{7}$$

and parameter of the other layers $\theta_t^*, t = 1, 2, \cdots, T$ *will maximize the Hamiltonian functions*

$$\sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*) \geq \sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta), \forall\theta \in \Theta_t \tag{8}$$

WHY DECOUPLING

Theorem 1. *(PMP for adversarial defense) There exists co-state processes* $p_s^* := p_{s,t}^* : t = 0, \cdots, T$ *such that the following holds for all* $t \in [T]$ *and* $s \in [S]$*:*

$$x_{s,t+1}^* = \nabla_p H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad x_{s,0}^* = x_{s,0} + \eta \tag{5}$$

$$p_{s,t}^* = \nabla_x H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*), \qquad p_{s,T}^* = -\frac{1}{S}\nabla\Phi(x_{s,T}^*) \tag{6}$$

At the same the the parameter of the first layer θ_0^* *satisfies*

$$\sum_{s=1}^{S} H_t(x_{s,0} + \hat{\eta}, p_{s,t+1}^*, \theta_0^*), \forall\theta \in \Theta_t \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta_0^*) \geq \sum_{s=1}^{S} H_0(x_{s,0}^*, p_{s,1}^*, \theta), \forall\theta \in \Theta_0, \|\hat{\eta}\|_\infty \leq \epsilon$$

Layer-wise maximal principle

Adversarial example exists in the first layer

and parameter of the other layers $\theta_t^*, t = 1, 2, \cdots, T$ *will maximize the Hamiltonian functions*

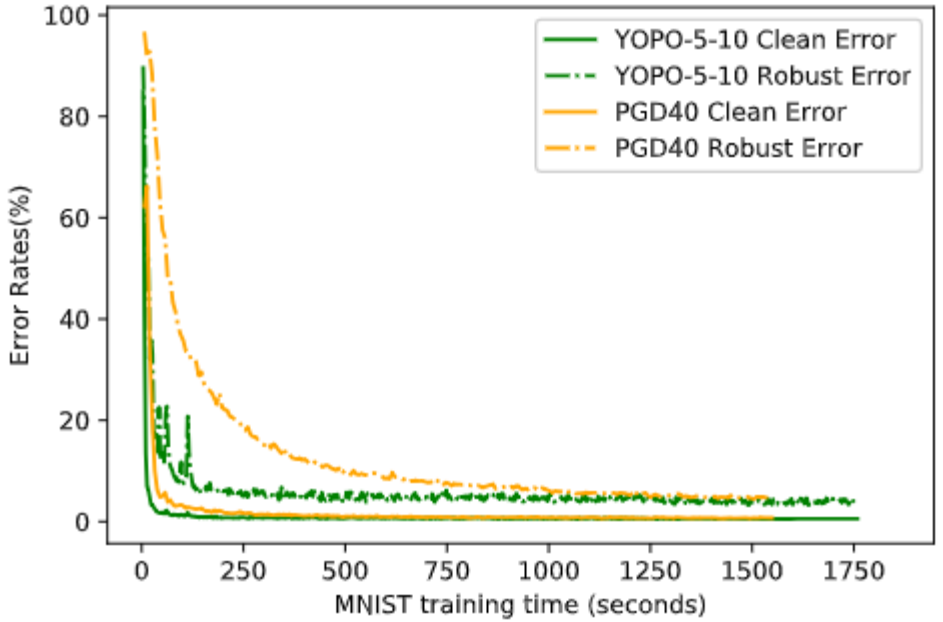
$$\sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta_t^*) \geq \sum_{s=1}^{S} H_t(x_{s,t}^*, p_{s,t+1}^*, \theta), \forall\theta \in \Theta_t \tag{8}$$

RESULT

Training Methods	Clean Data	PGD-20 Attack	Training Time (mins)
Natural train	95.03%	0.00%	233
PGD-3 [24]	90.07%	39.18%	1134
PGD-5 [24]	89.65%	43.85%	1574
PGD-10 [24]	87.30%	47.04%	2713
Free-8 [28][1]	86.29%	47.00%	667
YOPO-3-5 (Ours)	87.27%	43.04%	299
YOPO-5-3 (Ours)	86.70%	47.98%	**476**

[1] Code from https://github.com/ashafahi/free_adv_train.

Table 3: Results of Wide ResNet34 for CIFAR10.



(a) "Samll CNN" in[42] Result On MNIST