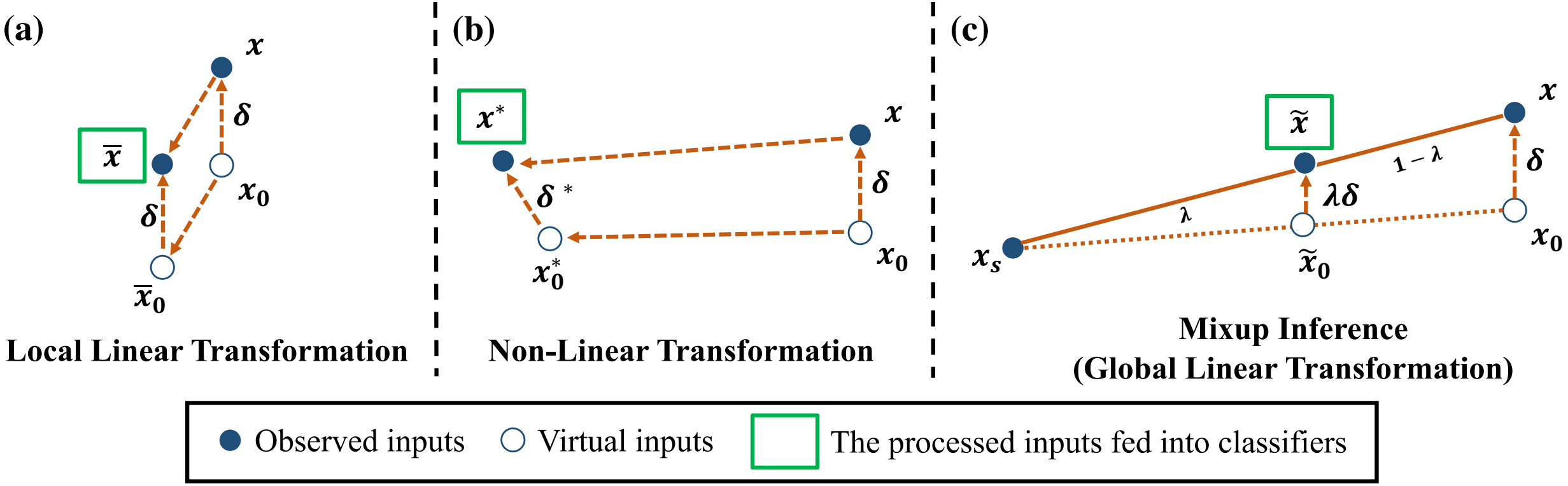
Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks

(ICLR 2020)

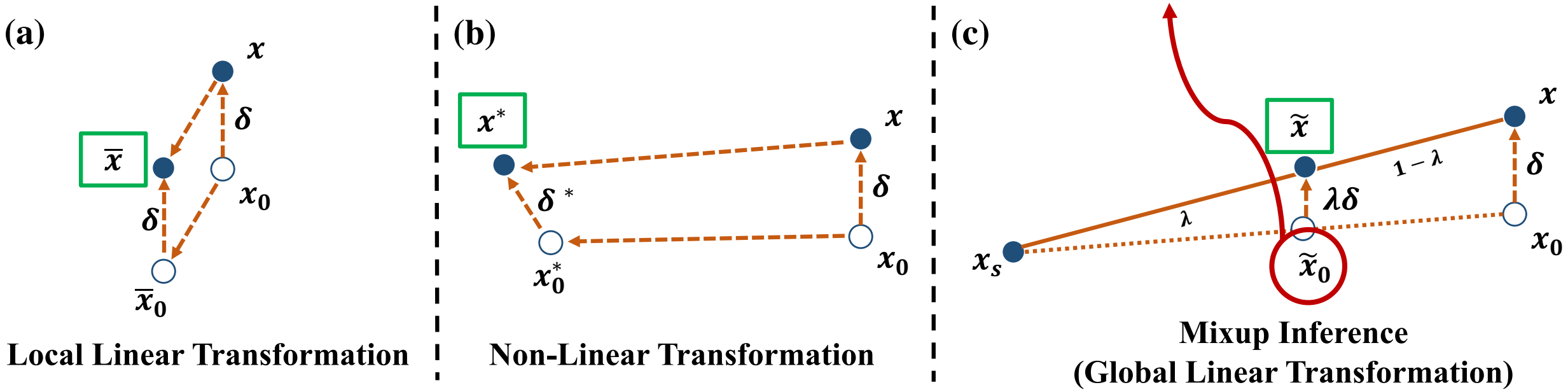
Tianyu Pang*, Kun Xu* and Jun Zhu

Motivation


(a)
x
δ
$\bar{x}$
x_0
δ
$\bar{x}_0$
Local Linear Transformation
(b)
x^*
x
δ^*
δ
x_0^*
x_0
Non-Linear Transformation
(c)
$\tilde{x}$
x
$1-\lambda$
δ
λ
$\lambda\delta$
x_s
$\tilde{x}_0$
x_0
Mixup Inference
(Global Linear Transformation)
Observed inputs
Virtual inputs
The processed inputs fed into classifiers

Motivation

the identity of x_0 can be recovered from the statistics of $\tilde{x}_0$ for mixup-trained models



Mixup-trained Models

We consider the setting of

$$\mathbf{1}_{y \neq \hat{y}} = \mathbf{1}_{z=1}.$$

Mixup-trained Models

We consider the setting of

$$\mathbf{1}_{y \neq \hat{y}} = \mathbf{1}_{z=1}.$$

Theoretically, with unlimited capability and sufficient clean samples, a well mixup-trained model F can be denoted as a linear function H on the convex combinations of clean examples (Hornik et al., 1989; Guo et al., 2019), i.e., $\forall x_i, x_j \sim p(x)$ and $\lambda \in [0, 1]$, there is

$$H(\lambda x_i + (1-\lambda)x_j) = \lambda H(x_i) + (1-\lambda)H(x_j). \tag{3}$$

mixup of two clean inputs

Mixup-trained Models

We consider the setting of

$$\mathbf{1}_{y \neq \hat{y}} = \mathbf{1}_{z=1}.$$

Theoretically, with unlimited capability and sufficient clean samples, a well mixup-trained model F can be denoted as a linear function H on the convex combinations of clean examples (Hornik et al., 1989; Guo et al., 2019), i.e., $\forall x_i, x_j \sim p(x)$ and $\lambda \in [0, 1]$, there is

$$H(\lambda x_i + (1-\lambda) x_j) = \lambda H(x_i) + (1-\lambda) H(x_j). \tag{3}$$

mixup of two clean inputs

$$F(x) = F(x_0 + \delta \cdot \mathbf{1}_{z=1}) = H(x_0) + G(\delta; x_0) \cdot \mathbf{1}_{z=1}. \tag{4}$$

potential perturbation

Mixup Inference (MI)

According to Eq. (3) and Eq. (4), the output of $\tilde{x}$ in MI is given by:

$$\begin{aligned} F(\tilde{x}) &= H(\tilde{x}_0) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1} \\ &= \lambda H(x_0) + (1-\lambda) H(x_s) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}, \end{aligned}$$

where $\tilde{x}_0 = \lambda x_0 + (1-\lambda) x_s$ is a virtual unperturbed counterpart of $\tilde{x}$ as shown in Fig. 1(c).

Mixup Inference (MI)

According to Eq. (3) and Eq. (4), the output of $\tilde{x}$ in MI is given by:

$$\begin{aligned} F(\tilde{x}) &= H(\tilde{x}_0) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1} \\ &= \lambda H(x_0) + (1-\lambda) H(x_s) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}, \end{aligned}$$

where $\tilde{x}_0 = \lambda x_0 + (1-\lambda) x_s$ is a virtual unperturbed counterpart of $\tilde{x}$ as shown in Fig. 1(c).

Note that there is $H_y(x_0) = 1$ and $H_{y_s}(x_s) = 1$, thus we have the y-th components as

$$\begin{aligned} F_y(x) &= 1 + G_y(\delta; x_0) \cdot \mathbf{1}_{z=1}; \\ F_y(\tilde{x}) &= \lambda + (1-\lambda) \cdot \mathbf{1}_{y=y_s} + G_y(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}. \end{aligned}$$

Mixup Inference (MI)

According to Eq. (3) and Eq. (4), the output of $\tilde{x}$ in MI is given by:

$$\begin{aligned} F(\tilde{x}) &= H(\tilde{x}_0) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1} \\ &= \lambda H(x_0) + (1-\lambda) H(x_s) + G(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}, \end{aligned}$$

where $\tilde{x}_0 = \lambda x_0 + (1-\lambda) x_s$ is a virtual unperturbed counterpart of $\tilde{x}$ as shown in Fig. 1(c).

Note that there is $H_y(x_0) = 1$ and $H_{y_s}(x_s) = 1$, thus we have the y-th components as

$$\begin{aligned} F_y(x) &= 1 + G_y(\delta; x_0) \cdot \mathbf{1}_{z=1}; \\ F_y(\tilde{x}) &= \lambda + (1-\lambda) \cdot \mathbf{1}_{y=y_s} + G_y(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}. \end{aligned}$$

Furthermore, according to Eq. (2), there is $\mathbf{1}_{y=\hat{y}} = \mathbf{1}_{z=0}$.

$$\begin{aligned} F_{\hat{y}}(x) &= \mathbf{1}_{z=0} + G_{\hat{y}}(\delta; x_0) \cdot \mathbf{1}_{z=1}; \\ F_{\hat{y}}(\tilde{x}) &= \lambda \cdot \mathbf{1}_{z=0} + (1-\lambda) \cdot \mathbf{1}_{\hat{y}=y_s} + G_{\hat{y}}(\lambda\delta; \tilde{x}_0) \cdot \mathbf{1}_{z=1}. \end{aligned}$$

Mixup Inference (MI)

When is the robustness improved?

$$\Delta F_{y}(x; p_s)|_{z=1} > 0; \Delta F_{\hat{y}}(x; p_s)|_{z=1} < 0.$$

Mixup Inference (MI)

When is the robustness improved?

$$\Delta F_{y}(x; p_s)|_{z=1} > 0; \Delta F_{\hat{y}}(x; p_s)|_{z=1} < 0.$$

MI with predicted label (MI-PL): In this case, the sampling label y_s is the same as the predicted label $\hat{y}$, i.e., $p_s(y) = \mathbf{1}_{y=\hat{y}}$ is a Dirac distribution on $\hat{y}$.

$$\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0) \right] \begin{cases} > 1 - \lambda, & \text{if } k = \hat{y}, \\ < \lambda - 1, & \text{if } k = y. \end{cases}$$

Mixup Inference (MI)

When is the robustness improved?

$$\Delta F_{y}(x;p_s)|_{z=1} > 0; \Delta F_{\hat{y}}(x;p_s)|_{z=1} < 0.$$

MI with predicted label (MI-PL): In this case, the sampling label y_s is the same as the predicted label $\hat{y}$, i.e., $p_s(y) = \mathbf{1}_{y=\hat{y}}$ is a Dirac distribution on $\hat{y}$.

$$\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0) \right] \begin{cases} > 1-\lambda, & \text{if } k = \hat{y}, \\ < \lambda - 1, & \text{if } k = y. \end{cases}$$

MI with other labels (MI-OL): In this case, the label y_s is uniformly sampled from the labels other than $\hat{y}$, i.e., $p_s(y) = \mathcal{U}_{\hat{y}}(y)$ is a discrete uniform distribution on the set $\{y \in [L] | y \neq \hat{y}\}$.

$$\mathbb{E}_{y_s \sim \mathcal{U}_{\hat{y}}(y)} \mathbb{E}_{x_s \sim p_s(x|y_s)} \left[G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0) \right] \begin{cases} > 0, & \text{if } k = \hat{y}, \\ < \frac{(\lambda-1)(L-2)}{L-1}, & \text{if } k = y. \end{cases}$$

Mixup Inference (MI)

When is the robustness improved?

$$\Delta F_y(x; p_s)|_{z=1} > 0; \Delta F_{\hat{y}}(x; p_s)|_{z=1} < 0.$$

MI with predicted label (MI-PL): In this case, the sampling label y_s is the same as the predicted label $\hat{y}$, i.e., $p_s(y) = 1_{y=\hat{y}}$ is a Dirac distribution on $\hat{y}$.

$$\mathbb{E}_{x_s \sim p_s(x|\hat{y})} [G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0)] \begin{cases} > 1 - \lambda, & \text{if } k = \hat{y}, \\ < \lambda - 1, & \text{if } k = y. \end{cases}$$

MI with other labels (MI-OL): In this case, the label y_s is uniformly sampled from the labels other than $\hat{y}$, i.e., $p_s(y) = \mathcal{U}_{\hat{y}}(y)$ is a discrete uniform distribution on the set $\{y \in [L] | y \neq \hat{y}\}$.

$$\mathbb{E}_{y_s \sim \mathcal{U}_{\hat{y}}(y)} \mathbb{E}_{x_s \sim p_s(x|y_s)} [G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0)] \begin{cases} > 0, & \text{if } k = \hat{y}, \\ < \frac{(\lambda-1)(L-2)}{L-1}, & \text{if } k = y. \end{cases}$$

Two Mechanisms of MI

$$\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0) \right]$$

$$\underbrace{\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\delta; \tilde{x}_0) \right]}_{\textbf{input transfer}} + \underbrace{\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; \tilde{x}_0) - G_k(\lambda\delta; \tilde{x}_0) \right]}_{\textbf{perturbation shrinkage}}.$$

Two Mechanisms of MI

$$\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\lambda\delta; \tilde{x}_0) \right]$$

$$\underbrace{\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; x_0) - G_k(\delta; \tilde{x}_0) \right]}_{\textbf{input transfer}} + \underbrace{\mathbb{E}_{x_s \sim p_s(x|\hat{y})} \left[G_k(\delta; \tilde{x}_0) - G_k(\lambda\delta; \tilde{x}_0) \right]}_{\textbf{perturbation shrinkage}}.$$

Obtain higher signal-to-noise ratio

Empirical Results

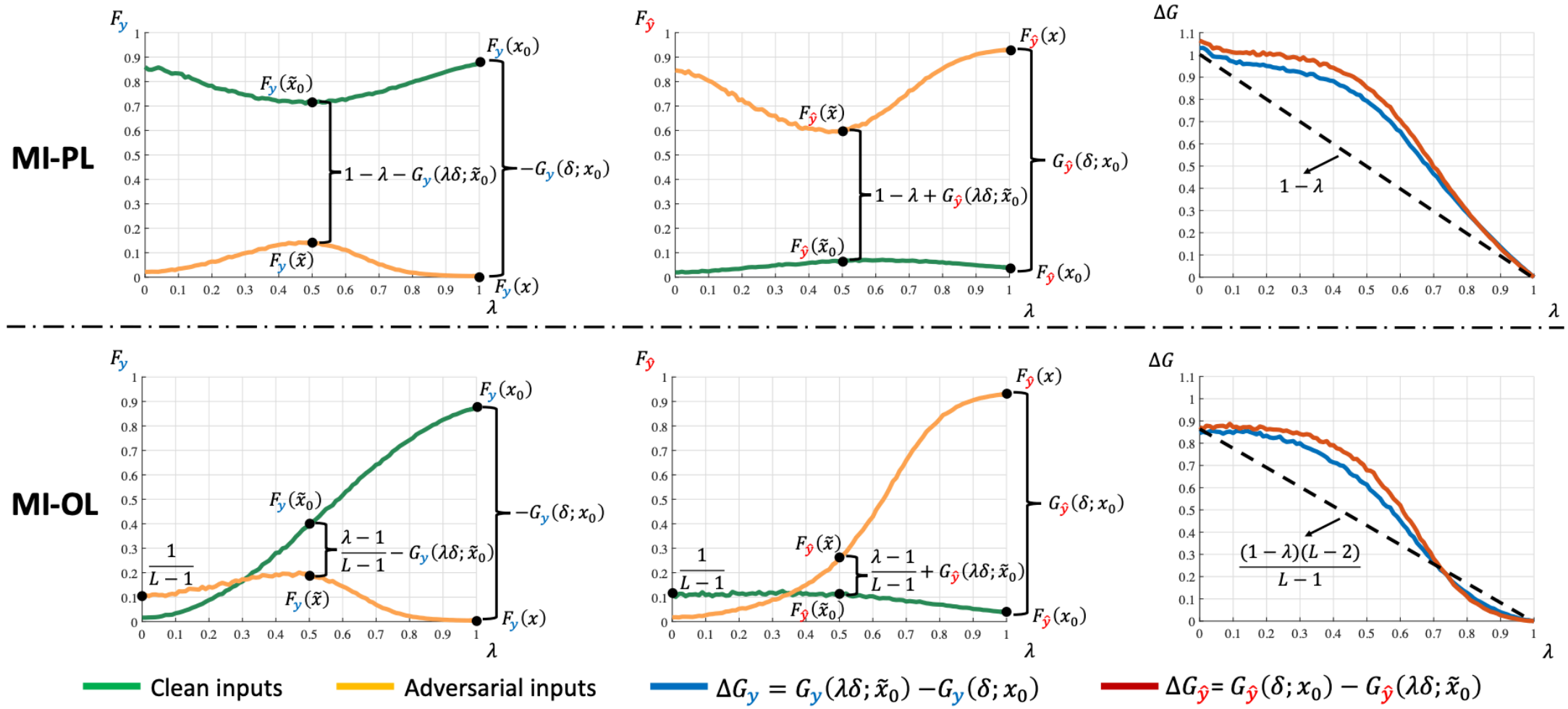
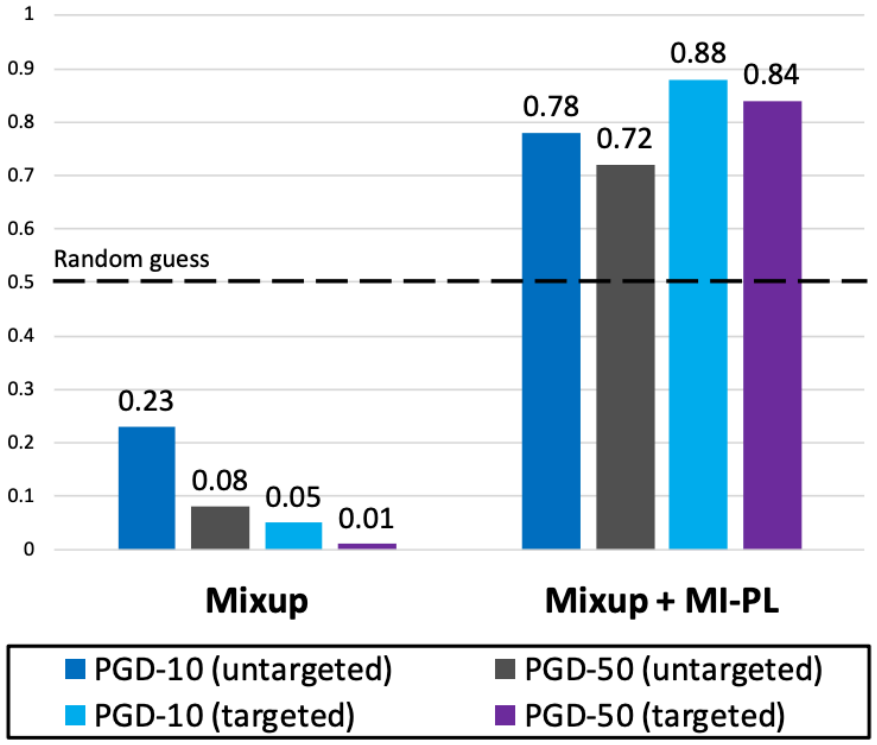


Figure 2: The results are averaged on 100 randomly test clean samples of **CIFAR-10**. The adversarial attack is untargeted PGD-10. Note that the ΔG_y calculated here is the minus value of it in Eq. (12) and Eq. (15).

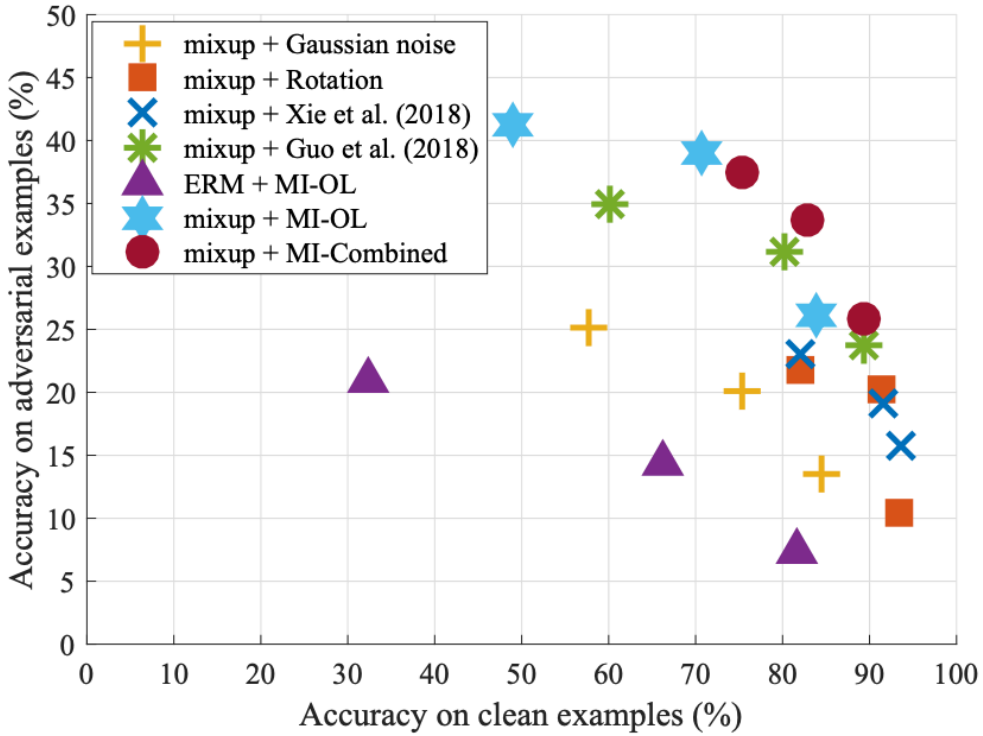
Empirical Results

Methods	Cle.	Untargeted Mode			Targeted Mode		
		PGD_{10}	PGD_{50}	PGD_{200}	PGD_{10}	PGD_{50}	PGD_{200}
Mixup	93.8	3.6	3.2	3.1	≤ 1	≤ 1	≤ 1
Mixup + Gaussian noise	84.4	13.5	9.6	8.8	37.7	28.6	27.9
Mixup + Random rotation	82.0	21.8	18.7	18.2	38.9	32.5	26.5
Mixup + Xie et al. (2018)	82.1	23.0	19.6	19.1	38.4	31.1	25.2
Mixup + Guo et al. (2018)	83.3	31.2	28.8	28.3	**57.8**	49.1	48.9
ERM + **MI-OL** (*ablation study*)	81.6	7.4	6.4	6.1	33.0	26.7	23.2
Mixup + **MI-OL**	83.9	26.1	18.8	18.3	55.6	**51.2**	**50.8**
Mixup + **MI-Combined**	82.9	**33.7**	**31.0**	**30.7**	56.1	49.7	49.4
Interpolated AT	89.7	46.7	43.5	42.5	65.6	62.5	61.9
Interpolated AT + Gaussian noise	84.7	55.6	53.7	53.5	70.1	69.1	69.0
Interpolated AT + Random rotation	83.4	57.8	56.7	55.9	69.8	68.2	67.4
Interpolated AT + Xie et al. (2018)	82.1	59.7	58.4	57.9	71.1	69.7	69.3
Interpolated AT + Guo et al. (2018)	83.9	60.9	60.7	60.3	73.2	72.1	71.6
AT + **MI-OL** (*ablation study*)	81.2	56.2	55.8	55.1	67.7	67.2	66.4
Interpolated AT + **MI-OL**	84.2	**64.5**	**63.8**	**63.3**	**75.3**	**74.7**	**74.5**

Empirical Results



(a) AUC scores



(b) Adversarial accuracy w.r.t clean accuracy

Figure 3: Results on **CIFAR-10**. **(a)** AUC scores on 1,000 randomly selected test clean samples and 1,000 adversarial counterparts crafted on these clean samples. **(b)** The adversarial accuracy w.r.t clean accuracy on 1,000 randomly selected test samples. The adversarial attack is untargeted PGD-10, with $\epsilon = 8/255$ and step size $2/255$. Each point for a certain method corresponds to a set of hyperparameters.

Empirical Results

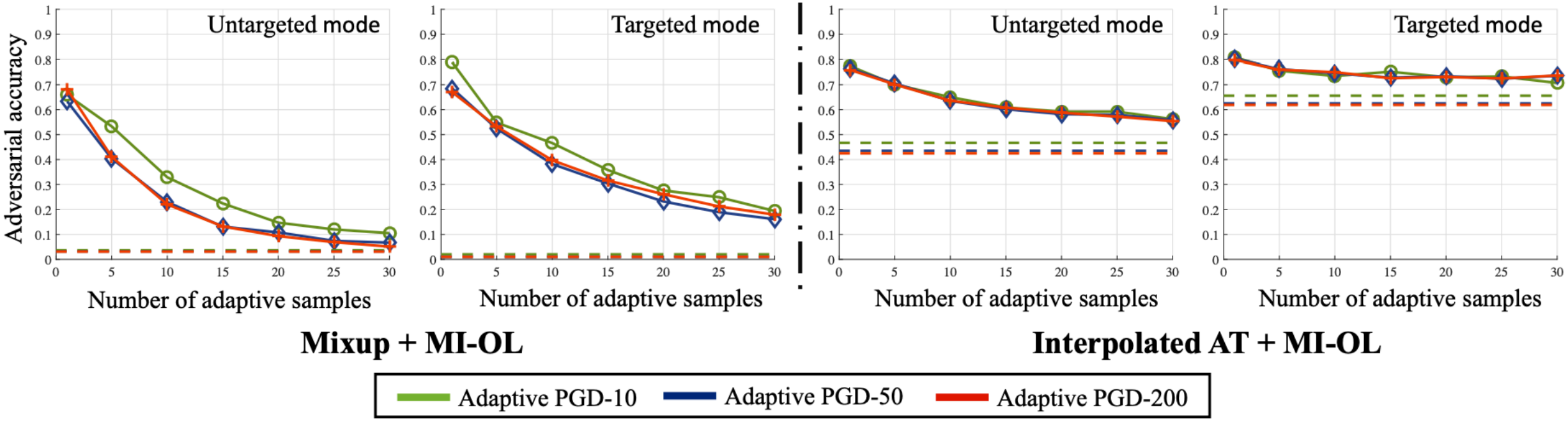


Figure 4: Classification accuracy under the adaptive PGD attacks on **CIFAR-10**. The number of adaptive samples refers to the execution times of sampling x_s in each iteration step of adaptive PGD. The dash lines are the accuracy of trained models without MI-OL under PGD attacks.

Thanks