Adversarial Training and Robustness for Multiple Perturbations

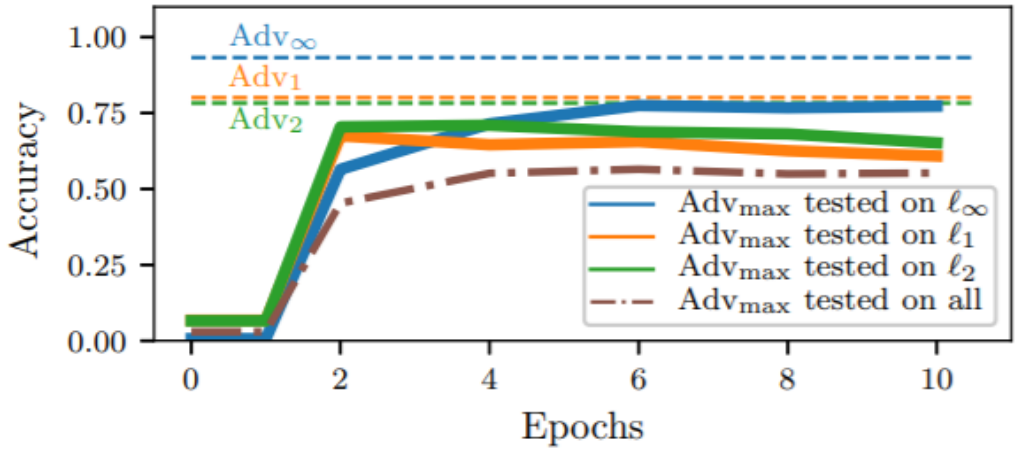
Florian Tramèr
Stanford University

Dan Boneh
Stanford University

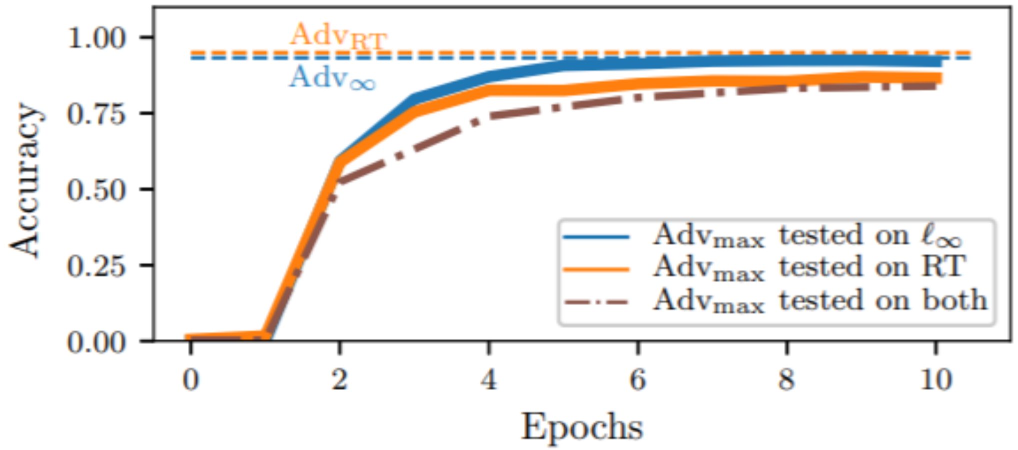
Hanqing Chao

Can we achieve adversarial robustness to different types of perturbations simultaneously?

For now, the answer is NO

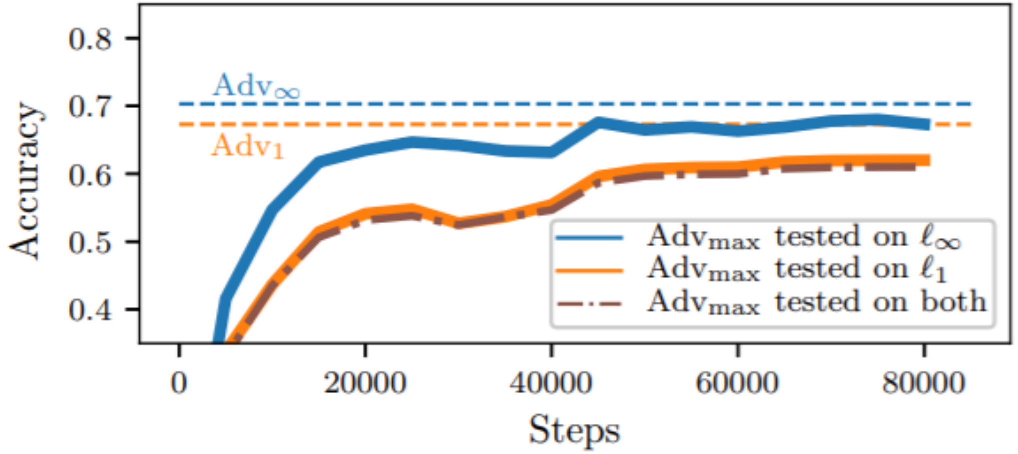


(a) MNIST models trained on ℓ_1, ℓ_2 & ℓ_∞ attacks.

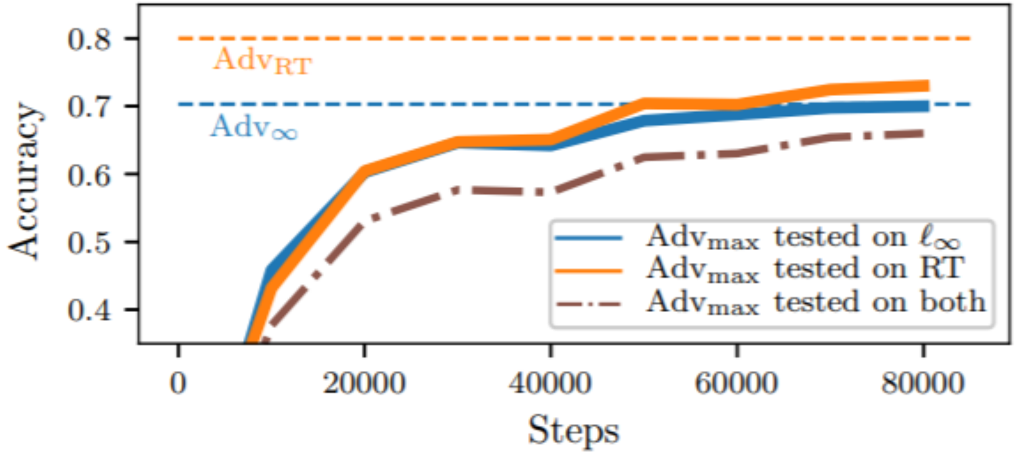


(b) MNIST models trained on ℓ_∞ and RT attacks.

RT: Rotation and translation



(c) CIFAR10 models trained on ℓ_1 and ℓ_∞ attacks.



(d) CIFAR10 models trained on ℓ_∞ and RT attacks.

Theoretical Proof

Notations:

- Data distribution $\mathcal{D}, (\boldsymbol{x}, y) \sim \mathcal{D}, \boldsymbol{x} \in \mathbb{R}^d, y \in [C]$
- Classifier $f: \mathbb{R}^d \longrightarrow [C]$
- Zero-one loss: $l(f(x), y) = \mathbb{1}_{f(\boldsymbol{x}) \neq y}$
- A type of perturbation: a set S, e.g., S can be an l_p-ball for a l_p perturbation

Theoretical Proof

Notations:

- Data distribution $\mathcal{D}$, $(\boldsymbol{x}, y) \sim \mathcal{D}$, $\boldsymbol{x} \in \mathbb{R}^d$, $y \in [C]$
- Classifier $f: \mathbb{R}^d \longrightarrow [C]$
- Zero-one loss: $l(f(x), y) = \mathbb{1}_{f(\boldsymbol{x}) \neq y}$
- A type of perturbation: a set S, e.g., S can be an l_p-ball for a l_p perturbation
- Adversarial risk: $\mathcal{R}_{adv}(f; S) := E_{(\boldsymbol{x},y)\sim\mathcal{D}}[\max_{\boldsymbol{r} \in S} l(f(\boldsymbol{x} + \boldsymbol{r}), y)]$
- For multiple perturbation sets $S_1, \ldots, S_n$:

$$\mathcal{R}_{\text{adv}}^{\max}(f; S_1, \ldots, S_n) := \mathcal{R}_{\text{adv}}(f; \cup_i S_i) , \quad \mathcal{R}_{\text{adv}}^{\text{avg}}(f; S_1, \ldots, S_n) := \frac{1}{n} \sum_i \mathcal{R}_{\text{adv}}(f; S_i) .$$

- Define S_1, S_2 are *Mutually Exclusive Perturbations* (MEPs), if $\mathcal{R}_{\text{adv}}^{\text{avg}}(f; S_1, S_2) \geq 1/|C|$

Theoretical Proof

l_∞ and l_1 perturbations are *Mutually Exclusive*

- Construct a specific distribution $\mathcal{D}$
- Prove that for $\forall\, f, \mathcal{R}_{\mathrm{adv}}^{\mathrm{avg}}(f; S_\infty, S_1) \geq 1/2$

Theoretical Proof

l_∞ and l_1 perturbations are *Mutually Exclusive*

- Construct a specific distribution $\mathcal{D}$
- Prove that for $\forall f, \mathcal{R}_{\text{adv}}^{\text{avg}}(f; S_\infty, S_1) \geq 1/2$

$$y \overset{u.a.r}{\sim} \{-1, +1\}, \quad x_0 = \begin{cases} +y, & \text{w.p. } p_0, \\ -y, & \text{w.p. } 1 - p_0 \end{cases}, \quad x_1, \ldots, x_d \overset{i.i.d}{\sim} \mathcal{N}(y\eta, 1), \tag{2}$$

where $p_0 \geq 0.5$, $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution and $\eta = \alpha/\sqrt{d}$ for some positive constant α.

Theoretical Proof

l_∞ and l_1 perturbations are *Mutually Exclusive*

- Construct a specific distribution $\mathcal{D}$
- Prove that for $\forall f, \mathcal{R}_{\text{adv}}^{\text{avg}}(f; S_\infty, S_1) \geq 1/2$

$$y \overset{u.a.r}{\sim} \{-1,+1\}, \quad x_0 = \begin{cases} +y, & \text{w.p. } p_0, \\ -y, & \text{w.p. } 1-p_0 \end{cases}, \quad x_1,\dots,x_d \overset{i.i.d}{\sim} \mathcal{N}(y\eta, 1)\,, \tag{2}$$

where $p_0 \geq 0.5$, $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution and $\eta = \alpha/\sqrt{d}$ for some positive constant α.

Theorem 1. *Let f be a classifier for $\mathcal{D}$. Let S_∞ be the set of ℓ_∞-bounded perturbations with $\epsilon = 2\eta$, and S_1 the set of ℓ_1-bounded perturbations with $\epsilon = 2$. Then, $\mathcal{R}_{adv}^{avg}(f; S_\infty, S_1) \geq 1/2$.*

Theoretical Proof

l_∞ and Spatial perturbations are (nearly) *Mutually Exclusive*

Theorem 2. *Let f be a classifier for $\mathcal{D}$ (with $x_0 \sim \mathcal{N}(y, \alpha^{-2})$). Let S_∞ be the set of ℓ_∞-bounded perturbations with $\epsilon = 2\eta$, and S_{RT} be the set of perturbations for an RT adversary with budget N. Then,* $\boxed{\mathcal{R}_{adv}^{avg}(f; S_\infty, S_{RT}) \geq 1/2 - O(1/\sqrt{N})}$.

As the distribution $\mathcal{D}$ is constructed, these two theorem might not hold in a real dataset.

Multi-perturbation ADV training

For a normal adversarial training: $L(f(\mathcal{A}(\boldsymbol{x})), y)$

where $\mathcal{A}(\cdot)$ is an attack procedure.

1. **"Max" strategy:** For each input $\boldsymbol{x}$, we train on the strongest adversarial example from all attacks, i.e., the max in $\hat{\mathcal{R}}_{\text{adv}}$ is replaced by $L(f(\mathcal{A}_{k^*}(\boldsymbol{x})), y)$, for $k^* = \arg\max_k L(f(\mathcal{A}_k(\boldsymbol{x})), y)$.

Multi-perturbation ADV training

For a normal adversarial training: $L(f(\mathcal{A}(\boldsymbol{x})), y)$

where $\mathcal{A}(\cdot)$ is an attack procedure.

1. **"Max" strategy:** For each input $\boldsymbol{x}$, we train on the strongest adversarial example from all attacks, i.e., the max in $\hat{\mathcal{R}}_{\text{adv}}$ is replaced by $\underline{L(f(\mathcal{A}_{k^*}(\boldsymbol{x})), y)\text{, for } k^* = \arg\max_k L(f(\mathcal{A}_k(\boldsymbol{x})), y)}$.
2. **"Avg" strategy:** This strategy simultaneously trains on adversarial examples from all attacks. That is, the max in $\hat{\mathcal{R}}_{\text{adv}}$ is replaced by $\underline{\frac{1}{n}\sum_{i=1}^{n} L(f(\mathcal{A}_i(\boldsymbol{x}), y))}$.

Affine Attacks

- $\beta \in [0,1]$
- $\boldsymbol{r} = \boldsymbol{r_1} + \boldsymbol{r_2}$
- $\boldsymbol{r_1} \in \beta \cdot S_1$
- $\boldsymbol{r_2} \in (1 - \beta) \cdot S_2$

Affine Attacks

- $\beta \in [0,1]$
- $\boldsymbol{r} = \boldsymbol{r_1} + \boldsymbol{r_2}$
- $\boldsymbol{r_1} \in \beta \cdot S_1$
- $\boldsymbol{r_2} \in (1-\beta) \cdot S_2$

An affine attack is equal or stronger than the union of multiple attacks

Affine Attacks

- $\beta \in [0,1]$
- $\boldsymbol{r} = \boldsymbol{r_1} + \boldsymbol{r_2}$
- $\boldsymbol{r_1} \in \beta \cdot S_1$
- $\boldsymbol{r_2} \in (1-\beta) \cdot S_2$

An affine attack is equal or stronger than the union of multiple attacks

Claim 3. *For a linear classifier $f(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x}+b)$, we have $\mathcal{R}_{adv}^{max}(f; S_p, S_q) = \mathcal{R}_{adv}(f; S_{affine})$.*

Affine Attacks

- $\beta \in [0,1]$
- $\boldsymbol{r} = \boldsymbol{r_1} + \boldsymbol{r_2}$
- $\boldsymbol{r_1} \in \beta \cdot S_1$
- $\boldsymbol{r_2} \in (1-\beta) \cdot S_2$

An affine attack is equal or stronger than the union of multiple attacks

Claim 3. *For a linear classifier $f(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x} + b)$, we have $\mathcal{R}_{adv}^{max}(f; S_p, S_q) = \mathcal{R}_{adv}(f; S_{affine})$.*

Theorem 4. *Let $f(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x} + b)$ be a linear classifier for $\mathcal{D}$ (with $x_0 \sim \mathcal{N}(y, \alpha^{-2})$). Let S_∞ be some ℓ_∞-ball and S_{RT} be rotation-translations with budget $N > 2$. Define S_{affine} as above. Assume $w_0 > w_i > 0, \forall i \in [1, d]$. Then $\mathcal{R}_{adv}(f; S_{affine}) > \mathcal{R}_{adv}^{max}(f; S_\infty, S_{RT})$.*

Experiments

Experiments on MNIST

Model	Acc.	ℓ_∞	ℓ_1	ℓ_2	$1-\mathcal{R}_{\text{adv}}^{\max}$	$1-\mathcal{R}_{\text{adv}}^{\text{avg}}$
Nat	**99.4**	0.0	12.4	8.5	0.0	7.0
Adv_∞	**99.1**	**91.1**	12.1	11.3	6.8	38.2
Adv_1	98.9	0.0	**78.5**	50.6	0.0	43.0
Adv_2	98.5	0.4	68.0	**71.8**	0.4	46.7
Adv_{avg}	97.3	76.7	53.9	58.3	**49.9**	**63.0**
$\text{Adv}_{\max}$	97.2	71.7	62.6	56.0	**52.4**	**63.4**

Model	Acc.	ℓ_∞	RT	$1-\mathcal{R}_{\text{adv}}^{\max}$	$1-\mathcal{R}_{\text{adv}}^{\text{avg}}$
Nat	**99.4**	0.0	0.0	0.0	0.0
Adv_∞	**99.1**	**91.4**	0.2	0.2	45.8
Adv_{RT}	**99.3**	0.0	**94.6**	0.0	47.3
Adv_{avg}	**99.2**	88.2	86.4	**82.9**	**87.3**
$\text{Adv}_{\max}$	98.9	89.6	85.6	**83.8**	**87.6**

Experiments

Experiments on CIFAR10

Model	Acc.	ℓ_∞	ℓ_1	$1-\mathcal{R}_{\text{adv}}^{\text{max}}$	$1-\mathcal{R}_{\text{adv}}^{\text{avg}}$
Nat	**95.7**	0.0	0.0	0.0	0.0
Adv_∞	92.0	**71.0**	16.4	16.4	44.9
Adv_1	90.8	53.4	**66.2**	53.1	60.0
Adv_{avg}	91.1	64.1	60.8	**59.4**	**62.5**
Adv_{max}	91.2	65.7	62.5	**61.1**	**64.1**

Model	Acc.	ℓ_∞	RT	$1-\mathcal{R}_{\text{adv}}^{\text{max}}$	$1-\mathcal{R}_{\text{adv}}^{\text{avg}}$
Nat	**95.7**	0.0	5.9	0.0	3.0
Adv_∞	92.0	**71.0**	8.9	8.7	40.0
Adv_{RT}	**94.9**	0.0	**82.5**	0.0	41.3
Adv_{avg}	93.6	67.8	78.2	**65.2**	**73.0**
Adv_{max}	93.1	**69.6**	75.2	**65.7**	**72.4**

Experiments

Experiments on affine attacks

Dataset	Attacks	acc. on S_{U}	acc. on S_{affine}
MNIST	ℓ_∞ & RT	83.8	62.6
CIFAR10	ℓ_∞ & RT	65.7	56.0
CIFAR10	ℓ_∞ & ℓ_1	61.1	58.0

Thanks！