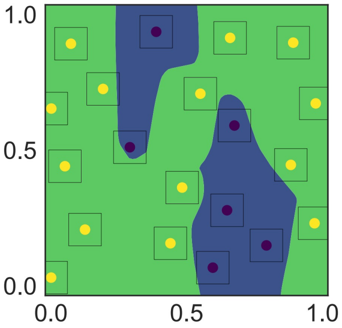
Theoretically Principled Trade-off between Robustness and Accuracy

Chao Chen

LEHIGH UNIVERSITY

Notations



A sample $\boldsymbol{x} \in \mathcal{X} \subseteq R^d$ and label $y \in \{+1, -1\}$,

$|\boldsymbol{x}|$ is the norm, e.g., $|\boldsymbol{x}|_\infty, |\boldsymbol{x}|_2$.

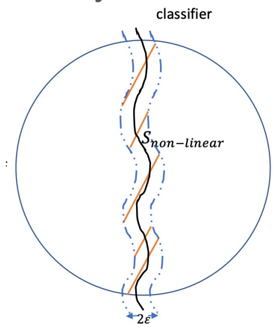
$\mathbb{B}(\boldsymbol{x}, \epsilon) = \{\boldsymbol{x}' \in \mathcal{X}\colon |\boldsymbol{x}' - \boldsymbol{x}| \leq \epsilon\}$ is the neighborhood of $\boldsymbol{x}$.

$f\colon \mathcal{X} \to \mathbb{R}$, a score function, maps an instance to a confidence value (being positive).

$sign\big(f(\cdot)\big)$ is the associated binary classifier, where $sign(\cdot)$ is the sign of input, and $sign(0) = 1$.

$\mathrm{DB}(f) = \{\boldsymbol{x} \in \mathcal{X}\colon f(\boldsymbol{x}) = 0\}$ is the decision boundary of f.

$\mathbb{B}(\mathrm{DB}(f), \epsilon) = \{\boldsymbol{x} \in \mathcal{X}\colon \exists \boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}) f(\boldsymbol{x}') \leq 0\}$ is the neighborhood of decision boundary.



For a given function $\psi(\boldsymbol{u})$, $\psi^*(\boldsymbol{v}) \coloneqq \sup_{\boldsymbol{u}}\{\boldsymbol{u}^T\boldsymbol{v} - \psi(\boldsymbol{u})\}$ is the conjugate function of ψ.

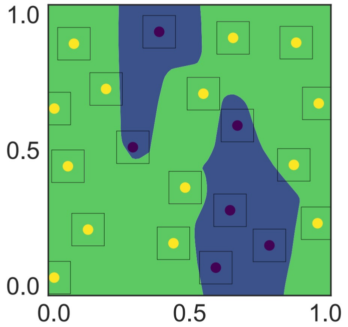
ψ^{**} is the bi-conjugate, and ψ^{-1} is the inverse function.

$1\{event\}$ is the indicator function indicating if $event$ happens.

Notations

$\mathbb{B}(\boldsymbol{x}, \epsilon) = \{\boldsymbol{x}' \in \mathcal{X}\colon |\boldsymbol{x}' - \boldsymbol{x}| \le \epsilon\}$ is the neighborhood of $\boldsymbol{x}$.

$\mathbb{B}(\mathrm{DB}(f), \epsilon) = \{\boldsymbol{x} \in \mathcal{X}\colon \exists \boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x}, \epsilon) \ \text{s.t.}\ f(\boldsymbol{x}) f(\boldsymbol{x}') \le 0\}$ is the neighborhood of decision boundary.



Assume that the data are drawn from an unknown distribution $(\boldsymbol{X}, Y) \sim \mathcal{D}$

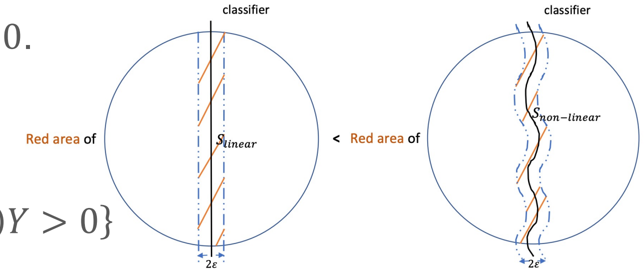
The robust (classification) error under ϵ perturbation:

$$\mathcal{R}_{\mathrm{rob}}(f) \coloneqq \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{\exists \boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon) \ \text{s.t.}\ f(\boldsymbol{X}')Y \le 0\}$$

The natural (classification) error:

$$\mathcal{R}_{\mathrm{nat}}(f) \coloneqq \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{f(\boldsymbol{X})Y \le 0\}$$

Clearly, $\mathcal{R}_{\mathrm{rob}}(f) \ge \mathcal{R}_{\mathrm{nat}}(f)$ for all f, and the equality holds when $\epsilon = 0$.



The boundary error:

$$\mathcal{R}_{\mathrm{bdy}}(f) \coloneqq \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0\}$$

And $\mathcal{R}_{\mathrm{rob}}(f) = \mathcal{R}_{\mathrm{nat}}(f) + \mathcal{R}_{\mathrm{bdy}}(f)$

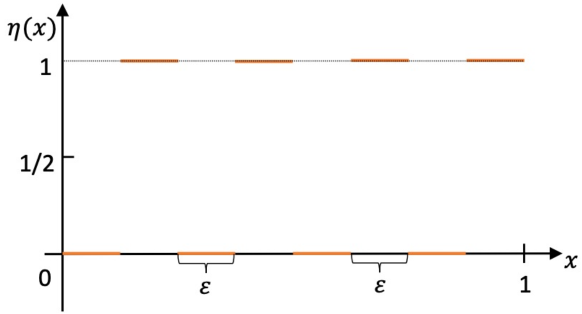
Toy example

The trade-off between natural and robust errors: training robust models may lead to a reduction of standard accuracy.

Assume that $\eta(x) := \Pr(Y = 1|X = x) = \begin{cases} 0, & x \in [2k\epsilon, (2k+1)\epsilon), \\ 1, & x \in \left((2k+1)\epsilon, (2k+1)\epsilon\right]. \end{cases}$ where $x \sim U[0,1]$

Bayes optimal classifier: $sign(2\eta(x) - 1)$

All-one classifier: 1 (always outputs "positive")



	Bayes Optimal Classifier	All-One Classifier
$\mathcal{R}_{\mathrm{nat}}$	0 (optimal)	1/2
$\mathcal{R}_{\mathrm{bdy}}$	1	0
$\mathcal{R}_{\mathrm{rob}}$	1	1/2 (optimal)

Toy example

LEHIGH UNIVERSITY

Assume that $\eta(x) := \Pr(Y = 1|X = x) = \begin{cases} 0, & x \in [2k\epsilon, (2k+1)\epsilon), \\ 1, & x \in ((2k+1)\epsilon, (2k+1)\epsilon]. \end{cases}$

The Bayes optimal classifier: $sign(2\eta(x) - 1)$

The all-one classifier: 1 (always outputs "positive")

	Bayes Optimal Classifier	All-One Classifier
$\mathcal{R}_{\text{nat}}$	0 (optimal)	1/2
$\mathcal{R}_{\text{bdy}}$	1	0
$\mathcal{R}_{\text{rob}}$	1	1/2 (optimal)

- For the natural error: $\mathcal{R}_{\text{nat}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{f(\boldsymbol{X})Y \le 0\}$:

It is obvious that $\mathcal{R}_{\text{nat}}(f) = 0$ for Bayes classifier, and $\mathcal{R}_{\text{nat}}(f) = 1/2$ for all-one classifier.

- For the boundary error $\mathcal{R}_{\text{bdy}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{\boldsymbol{X} \in \mathbb{B}(\text{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0\}$:

For Bayes classifier, we can always find a perturbation resulting in the right prediction, since the interval is ϵ.

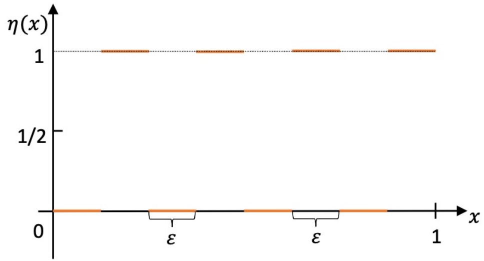
For all-one classifier, $\text{DB}(f)$ (if any) is not within [0,1], and thus the event never happens.

- For the robust error $\mathcal{R}_{\text{rob}}(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}} 1\{\exists \boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon) \text{ s.t. } f(\boldsymbol{X}')Y \le 0\}$:

For Bayes classifier, we can always find a perturbation to flip the prediction, since the interval is ϵ.

For all-one classifier, since $f(X) = 1, \forall X$, we have $1/2$ change to obtain negative sample ($Y = -1$).

Or we can compute it by $\mathcal{R}_{\text{rob}}(f) = \mathcal{R}_{\text{nat}}(f) + \mathcal{R}_{\text{bdy}}(f)$.



In most of existing works, we can assign different weights on both errors ($\mathcal{R}_{\text{nat}} + \mathcal{R}_{\text{bdy}}$) to balance them.

In this paper, the authors try to devise tight differentiable upper bounds on both terms, as both involve 0-1 loss functions.

Classification-calibrated surrogate loss

0-1 loss function is intractable -> tractable surrogate loss $\mathcal{R}_\phi(f) \coloneqq \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{D}}\phi(f(\boldsymbol{X})Y)$.

Define conditional ϕ-risk:

For $\eta \in [0,1]$, $H(\eta) \coloneqq \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha) \coloneqq \inf_{\alpha\in\mathbb{R}} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big)$,

and define $H^-(\eta) \coloneqq \inf_{\alpha:\alpha(2\eta-1)\le 0} C_\eta(\alpha) \coloneqq \inf_{\alpha:\alpha(2\eta-1)\le 0} \big(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\big)$.

Assumption on ϕ: it is classification-calibrated: if $H^-(\eta) > H(\eta)$ for any $\eta \neq 1/2$.

Intuition:

$\eta(x) \coloneqq \Pr(Y=1|X=x)$ and α is the probability of positive class predicted by f.

$H(\eta) = \min_f \mathcal{R}_{\text{nat}}(f)$,

$H^-(\eta) = \min_f \mathcal{R}_{\text{nat}}(f)$, s.t. f is inconsistent with Bayes optimal classifier

Classification-calibrated surrogate loss

The functional transform of classification-calibrated loss ϕ:

Define $\tilde{\psi}(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$ and $\psi: [0,1] \to [0, \infty)$ by $\psi = \tilde{\psi}^{**}$. (ψ^{*} is the conjugate function of ψ).

$\psi(\theta)$ is the largest convex lower bound on $\tilde{\psi}(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$

$\tilde{\psi}(\theta)$ characterizes how close the surrogate loss ϕ is to the class of non-classification-calibrated losses.

Property of classification-calibrated loss:

For classification-calibrated surrogate loss ϕ, ψ is non-decreasing, continuous, convex on $[0,1]$ and $\psi(0) = 0$.

Surrogate loss and 0-1 loss

Upper bound:

Let $\mathcal{R}_\phi(f) \coloneqq \mathbb{E}\phi(f(\boldsymbol{X})Y)$ and $\mathcal{R}_\phi^*(f)$, for non-negative classification-calibrated loss ϕ with $\phi(0) \geq 1$, any measurable $f: \mathcal{X} \to \mathbb{R}$, any probability distribution on $\mathcal{X}\times\{\pm 1\}$, and any $\lambda > 0$, we have:

$$\mathcal{R}_{\mathrm{rob}}(f) - \mathcal{R}_{\mathrm{nat}}^* \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \Pr[\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0]$$

$$\leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda)$$

The models are vulnerable to small adversarial attacks because the probability that data lie around the decision boundary of the model is large.

Surrogate loss and 0-1 loss

LEHIGH UNIVERSITY

Let $\mathcal{R}_\phi(f) \coloneqq \mathbb{E}\phi(f(\boldsymbol{X})Y)$ and $\mathcal{R}_\phi^*(f)$, for non-negative classification-calibrated loss ϕ with $\phi(0) \geq 1$, any measurable $f: \mathcal{X} \to \mathbb{R}$, any probability distribution on $\mathcal{X} \times \{\pm 1\}$, and any $\lambda > 0$, we have:

$$\mathcal{R}_{\mathrm{rob}}(f) - \mathcal{R}_{\mathrm{nat}}^* \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \Pr[\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0]$$

$$\leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda)$$

Proof:

The first inequality holds since ϕ is a classification-calibrated loss[1] and $\mathcal{R}_{\mathrm{bdy}} = \Pr[\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0]$:

$$\mathcal{R}_{\mathrm{rob}}(f) = \mathcal{R}_{\mathrm{nat}}(f) + \mathcal{R}_{\mathrm{bdy}}(f)$$

$$\mathcal{R}_{\mathrm{rob}}(f) - \mathcal{R}_{\mathrm{nat}}^* = \mathcal{R}_{\mathrm{nat}}(f) - \mathcal{R}_{\mathrm{nat}}^* + \mathcal{R}_{\mathrm{bdy}}(f) \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \mathcal{R}_{\mathrm{bdy}}(f)$$

Now we consider the second inequality:

$$\Pr[\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), f(\boldsymbol{X})Y > 0] \leq \Pr[\boldsymbol{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)]$$

$$= \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} 1\{f(\boldsymbol{X}') \neq f(\boldsymbol{X})\}$$

$$= \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} 1\{f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda < 0\}$$

$$\leq \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X}, \epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda)$$

[1] Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. "Convexity, classification, and risk bounds." Journal of the American Statistical Association 2006.

Surrogate loss and 0-1 loss

Lower bound:

Suppose that $|\mathcal{X}| \geq 2$. For non-negative classification-calibrated loss ϕ with $\phi(x) \to 0$ as $x \to +\infty$, and any $\xi > 0$, any $\theta \in [0,1]$. There exists a probability distribution on $\mathcal{X} \times \{\pm 1\}$, a function $f: \mathbb{R}^d \to \mathbb{R}$ and a regularization $\lambda > 0$ such that $\mathcal{R}_{\text{rob}}(f) - \mathcal{R}_{\text{nat}}^* = \theta$ and:

$$\psi\left(\theta - \mathbb{E} \max_{X' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(X')f(X)/\lambda)\right) \leq \mathcal{R}_\phi(f) - \mathcal{R}_\phi^* \leq \psi\left(\theta - \mathbb{E} \max_{X' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(X')f(X)/\lambda)\right) + \xi$$

Under the extra conditions on loss functions $\lim_{x \to +\infty} \phi(x) = 0$, the upper bound is tight.

The first inequality holds since ψ is non-decreasing, continuous, convex on $[0,1]$ and

$$\mathcal{R}_{\text{rob}}(f) - \mathcal{R}_{\text{nat}}^* \leq \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda)$$

Adversarial training by TRADES

Based on previous theorems, we consider a new surrogate loss:

$$\min_f \mathbb{E}\left\{\phi(f(\boldsymbol{X})Y) + \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X})f(\boldsymbol{X}')/\lambda)\right\}$$

The first term, $\phi(f(\boldsymbol{X})Y)$, minimizes the natural error.

The second regularization term, $\max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X})f(\boldsymbol{X}')/\lambda)$, minimizes the difference between the predictions of natural example and the adversarial example. Thus, it stands for the "robustness".

λ can balance the importance of natural and robust errors.

(It tends to be Bayes optimal classifier when $\lambda \to +\infty$ and all-one classifier when $\lambda \to 0$.)

We can easily extend it to multi-class tasks by replacing ϕ with a multi-class calibrated loss $\mathcal{L}(\cdot,\cdot)$:

$$\min_f \mathbb{E}\left\{\mathcal{L}(f(\boldsymbol{X}), Y) + \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \mathcal{L}\big(f(\boldsymbol{X}), f(\boldsymbol{X}')\big)/\lambda\right\}$$

In most of existing works:

$$\min_f \mathbb{E}\left\{\max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X}')Y)\right\}$$

is served as the upper bound of $\mathcal{R}_{\text{rob}}(f)$. However, it may not be the ***tight*** upper bound and may not capture the trade-off between natural and robust errors.

Adversarial training by TRADES

LEHIGH UNIVERSITY

Line 5: x_i is global minimizer to $g(x') := \mathcal{L}\big(f(x_i), f(x')\big)$, thus, initialize x_i' by adding small perturbation.

Line 7: solve $\max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \mathcal{L}\big(f(\boldsymbol{X}), f(\boldsymbol{X}')\big)/\lambda$ by projected gradient descent.

Line 10: gradient descent for the objective function

$$\min_f \mathbb{E}\left\{\mathcal{L}(f(\boldsymbol{X}), Y) + \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \mathcal{L}\big(f(\boldsymbol{X}), f(\boldsymbol{X}')\big)/\lambda\right\}$$

Algorithm 1 Adversarial training by TRADES

input Step sizes η_1 and η_2, batch size m, number of iterations K in inner optimization, network architecture parametrized by θ

output Robust network f_θ

1: Randomly initialize network f_θ, or initialize network with pre-trained configuration
2: **repeat**
3: Read mini-batch $B = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ from training set
4: **for** $i = 1, ..., m$ (in parallel) **do**
5: $\boldsymbol{x}_i' \leftarrow \boldsymbol{x}_i + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian distribution with zero mean and identity variance
6: **for** $k = 1, ..., K$ **do**
7: $\boldsymbol{x}_i' \leftarrow \Pi_{\mathbb{B}(\boldsymbol{x}_i,\epsilon)}(\eta_1 \mathsf{sign}(\nabla_{\boldsymbol{x}_i'} \mathcal{L}(f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_i'))) + \boldsymbol{x}_i')$, where Π is the projection operator
8: **end for**
9: **end for**
10: $\theta \leftarrow \theta - \eta_2 \sum_{i=1}^m \nabla_\theta [\mathcal{L}(f_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i) + \mathcal{L}(f_\theta(\boldsymbol{x}_i), f_\theta(\boldsymbol{x}_i'))/\lambda]/m$
11: **until** training converged

Experiments

Verify the tightness of upper bound.

$$\Delta_{LHS} = \mathcal{R}_{\text{rob}}(f) - \mathcal{R}_{\text{nat}}^* \le \psi^{-1}\left(\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right) + \mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda) = \Delta_{RHS}$$

Train a classifier with natural training method to estimate $\mathcal{R}_{\text{nat}}^* = 0\%$ and $\mathcal{R}_\phi^* = 0.0$

Find the classifier f by $\min_f \mathbb{E}\left\{\phi(f(\boldsymbol{X})Y) + \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X})f(\boldsymbol{X}')/\lambda)\right\}$ and approximate $\mathcal{R}_{\text{rob}}$ and $\mathcal{R}_\phi$.

Estimate $\mathbb{E} \max_{\boldsymbol{X}' \in \mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X}')f(\boldsymbol{X})/\lambda)$ by FGSM.

(The expectation is estimated in the test set.)

λ	$\mathcal{A}_{\text{rob}}(f)$ (%)	$\mathcal{R}_\phi(f)$	$\Delta = \Delta_{\text{RHS}} - \Delta_{\text{LHS}}$
2.0	99.43	0.0006728	0.006708
3.0	99.41	0.0004067	0.005914
4.0	99.37	0.0003746	0.006757
5.0	99.34	0.0003430	0.005860

Experiments

Robust accuracy $\mathcal{A}_{\text{rob}}(f) = 1 - \mathcal{R}_{\text{rob}}(f)$, and $\mathcal{A}_{\text{nat}}(f) = 1 - \mathcal{R}_{\text{nat}}(f)$

Sensitivity of λ

$1/\lambda$	$\mathcal{A}_{\text{rob}}(f)$ (%) on MNIST	$\mathcal{A}_{\text{nat}}(f)$ (%) on MNIST	$\mathcal{A}_{\text{rob}}(f)$ (%) on CIFAR10	$\mathcal{A}_{\text{nat}}(f)$ (%) on CIFAR10
1.0	94.75 ± 0.0712	99.28 ± 0.0125	44.68 ± 0.3088	87.01 ± 0.2819
2.0	95.45 ± 0.0883	99.29 ± 0.0262	48.22 ± 0.0740	85.22 ± 0.0543
3.0	95.57 ± 0.0262	99.24 ± 0.0216	49.67 ± 0.3179	83.82 ± 0.4050
4.0	95.65 ± 0.0340	99.16 ± 0.0205	50.25 ± 0.1883	82.90 ± 0.2217
5.0	95.65 ± 0.1851	99.16 ± 0.0403	50.64 ± 0.3336	81.72 ± 0.0286

Experiments

$$\min_f \mathbb{E}\left\{\max_{X' \in \mathbb{B}(X,\epsilon)} \phi(f(X')Y)\right\}$$

Defense	Defense type	Under which attack	Dataset	Distance	$\mathcal{A}_{\text{nat}}(f)$	$\mathcal{A}_{\text{rob}}(f)$
Buckman et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	0%
Ma et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	5%
Dhillon et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	0%
Song et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	9%
Na et al. (2017)	gradient mask	Athalye et al. (2018)	CIFAR10	0.015 (ℓ_∞)	-	15%
Wong et al. (2018)	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	27.07%	23.54%
Madry et al. (2018)	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	87.30%	**47.04%**
Zheng et al. (2016)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	94.64%	0.15%
Kurakin et al. (2017)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	85.25%	45.89%
Ross & Doshi-Velez (2017)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	95.34%	0%
TRADES ($1/\lambda = 1.0$)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	88.64%	49.14%
TRADES ($1/\lambda = 6.0$)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	84.92%	**56.61%**
TRADES ($1/\lambda = 1.0$)	regularization	DeepFool (ℓ_∞)	CIFAR10	0.031 (ℓ_∞)	88.64%	59.10%
TRADES ($1/\lambda = 6.0$)	regularization	DeepFool (ℓ_∞)	CIFAR10	0.031 (ℓ_∞)	84.92%	61.38%
TRADES ($1/\lambda = 1.0$)	regularization	LBFGSAttack	CIFAR10	0.031 (ℓ_∞)	88.64%	84.41%
TRADES ($1/\lambda = 6.0$)	regularization	LBFGSAttack	CIFAR10	0.031 (ℓ_∞)	84.92%	81.58%
TRADES ($1/\lambda = 1.0$)	regularization	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	88.64%	51.26%
TRADES ($1/\lambda = 6.0$)	regularization	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	84.92%	57.95%
TRADES ($1/\lambda = 1.0$)	regularization	C&W	CIFAR10	0.031 (ℓ_∞)	88.64%	84.03%
TRADES ($1/\lambda = 6.0$)	regularization	C&W	CIFAR10	0.031 (ℓ_∞)	84.92%	81.24%
Samangouei et al. (2018)	gradient mask	Athalye et al. (2018)	MNIST	0.005 (ℓ_2)	-	55%
Madry et al. (2018)	robust opt.	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.36%	96.01%
TRADES ($1/\lambda = 6.0$)	regularization	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.48%	96.07%
TRADES ($1/\lambda = 6.0$)	regularization	C&W	MNIST	0.005 (ℓ_2)	99.48%	99.46%

Thank you

LEHIGH UNIVERSITY