Defense Against Adversarial Attacks (Theoretic)

Recall the empirical defense approaches

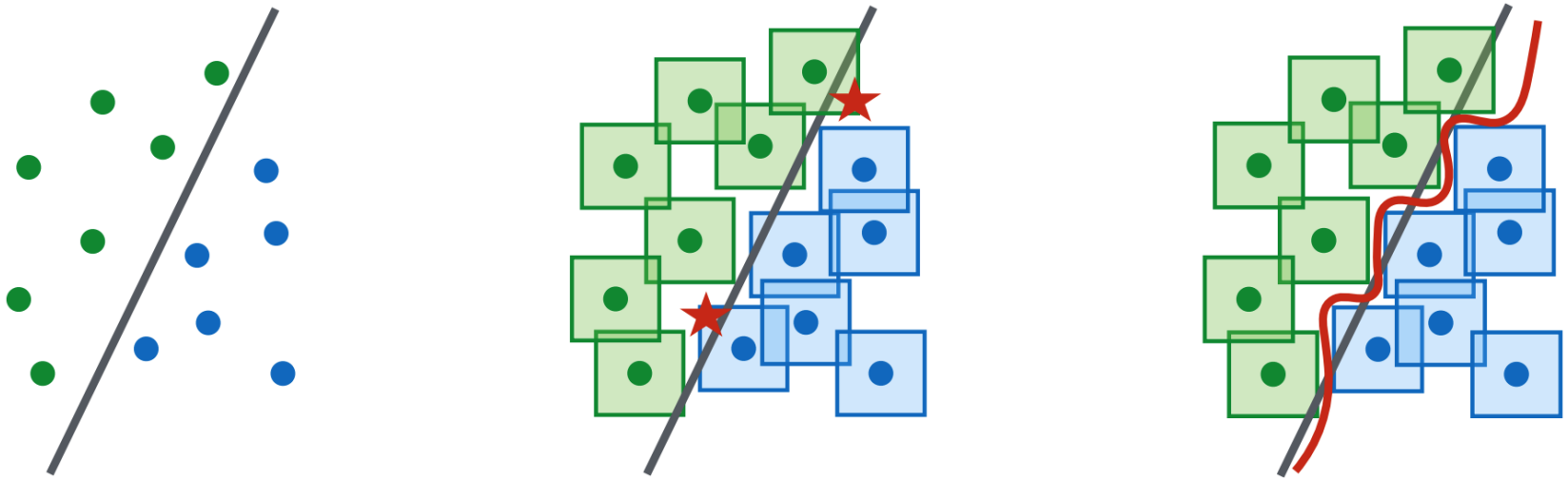
- PeerNet: leveraging the peer information (consistency)
- Distillation as a defense: ensure the classification output by a DNN remains constant in a closed neighborhood around any given sample extracted from the input distribution $\rho_{adv}(F) = E_{\mu}[\Delta_{adv}(X, F)]$
- PGD adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

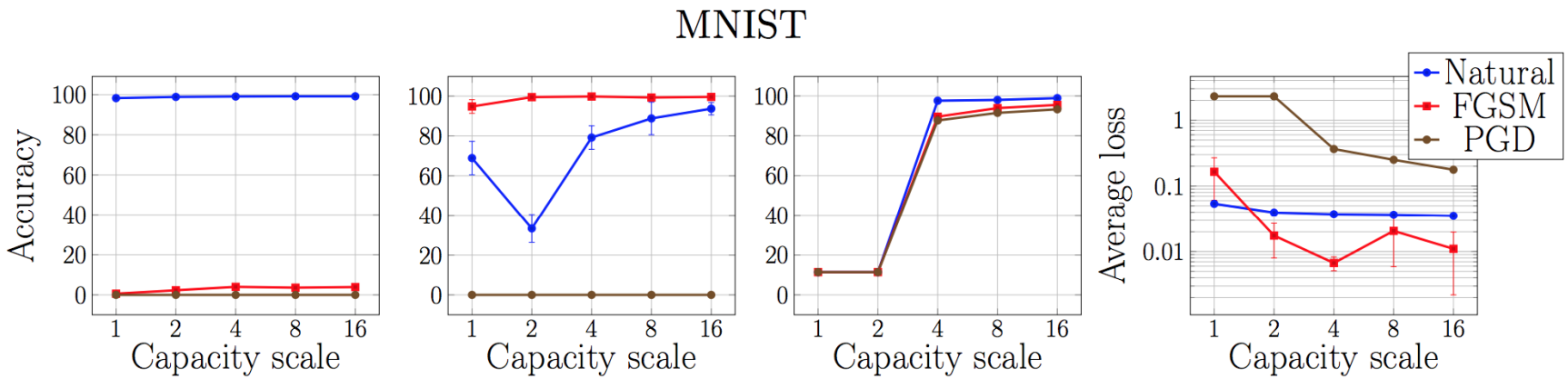
$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

- Use a natural saddle point (min-max) formulation to capture the notion of security against adversarial attacks in a principled manner.
- The formulation casts both attacks and defenses into a common theoretical framework.
- Motivate projected gradient descent (PGD) as a universal “first-order adversary”.

Model Capacity



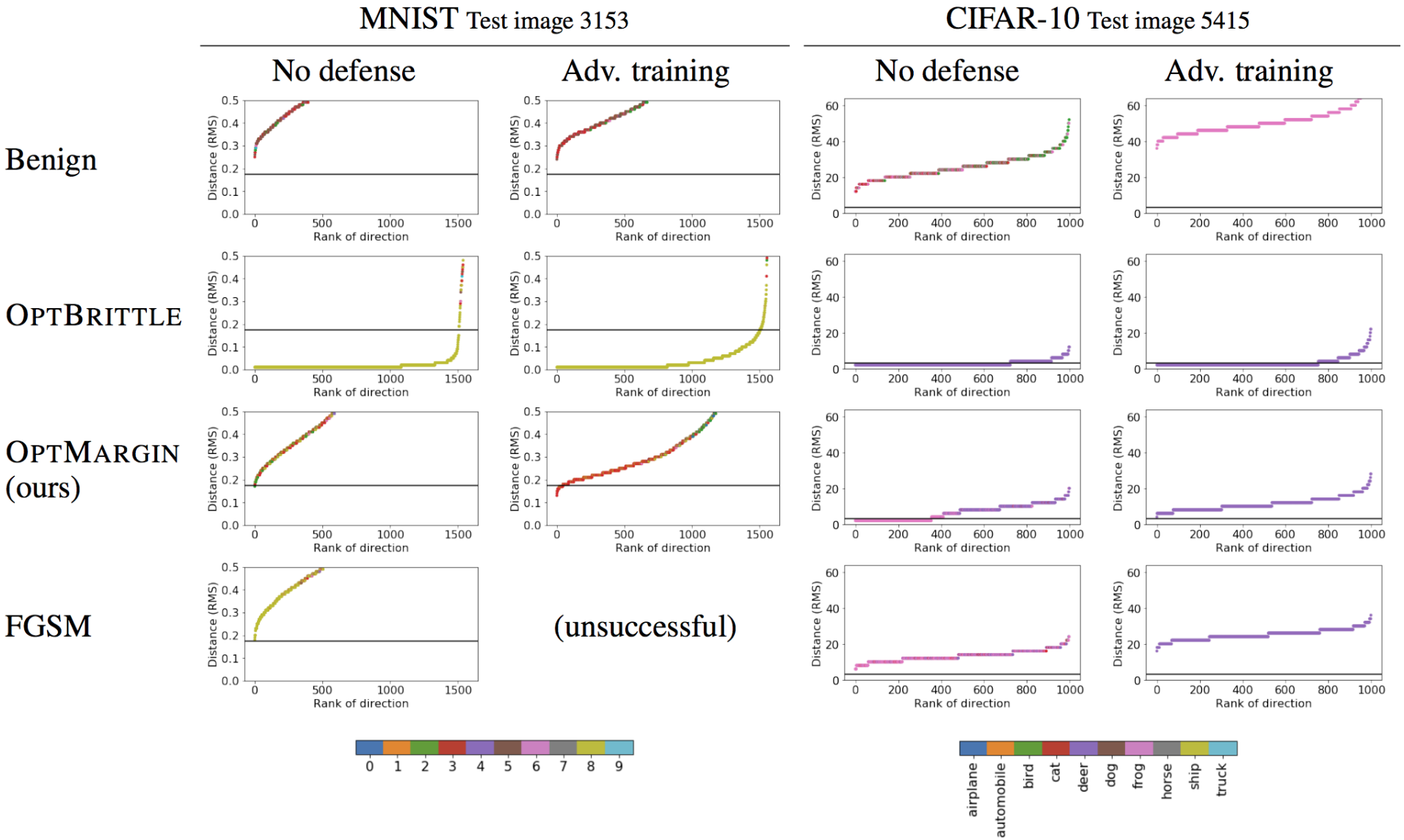
Towards Deep Learning Models Resistant to Adversarial Attacks



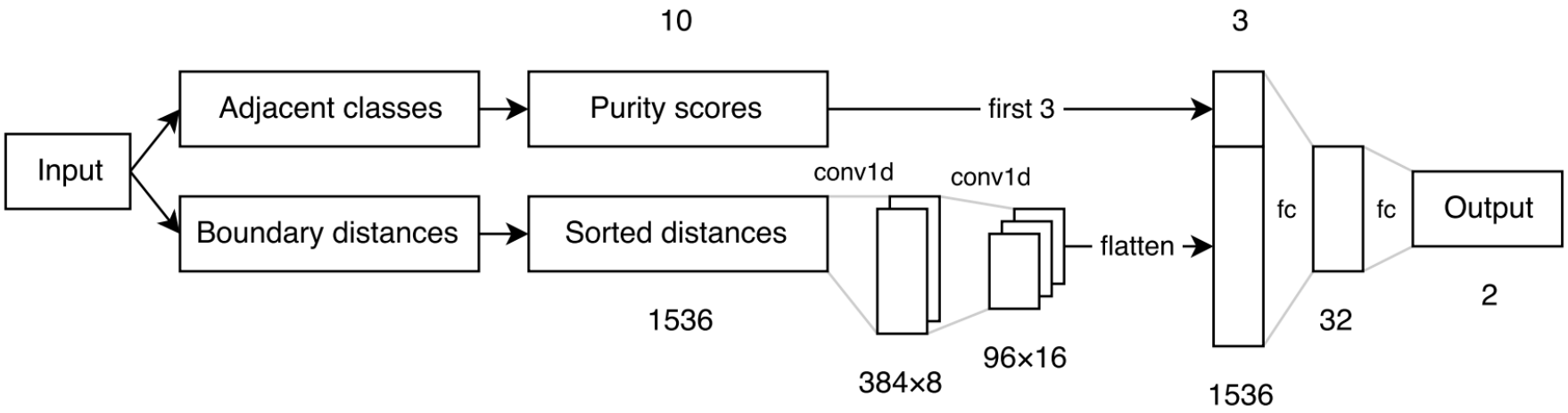
Beyond the Min-max Game

- Will it help if we have more knowledge about our learning tasks?
  - General understanding about ML models
  - Properties of specific learning tasks

Decision Boundary Based Detection



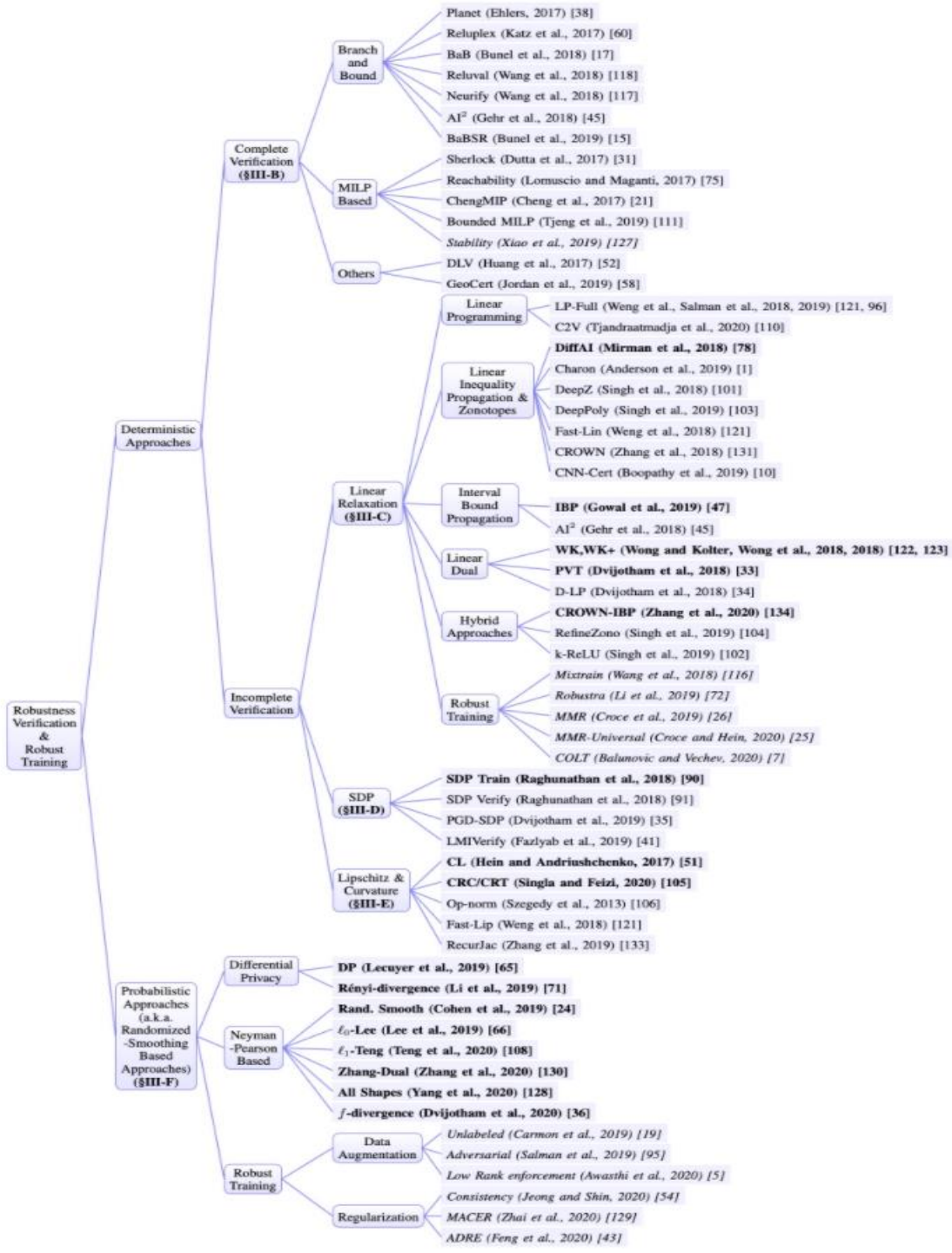
Decision Boundary Analysis of Adversarial Examples



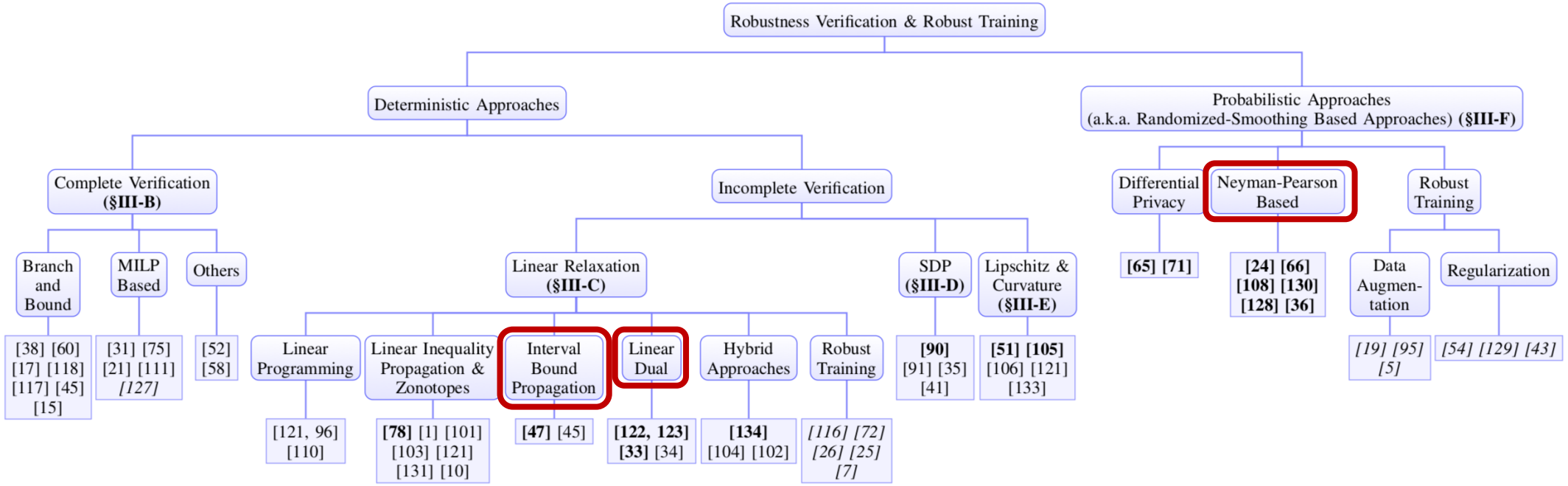
Training attack	False pos. Benign	False neg. OptBrittle	False neg. OptMargin	Accuracy Our approach	Accuracy Cao & Gong
MNIST, normal training					
OptBrittle	1.0%	1.0%	74.1%	90.4%	10%
OptMargin	**9.6%**	0.6%	7.2%		
MNIST, PGD adversarial training					
OptBrittle	2.6%	2.0%	39.8%		
OptMargin	10.3%	0.4%	14.5%		
CIFAR-10, normal training					
OptBrittle	5.3%	3.2%	56.8%	96.4%	5%
OptMargin	8.4%	7.4%	5.3%		
CIFAR-10, PGD adversarial training					
OptBrittle	0.0%	2.4%	51.8%		
OptMargin	**3.6%**	0.0%	1.2%		

Takeaways

- Decision boundaries of DNNs are important towards improving learning robustness
- Isolated islands in the data manifold would lead to harder detected/defensed adversarial behaviors


Robustness Verification & Robust Training
Deterministic Approaches
Probabilistic Approaches (a.k.a. Randomized -Smoothing Based Approaches) (§III-F)
Complete Verification (§III-B)
Incomplete Verification
Branch and Bound
Planet (Ehlers, 2017) [38]
Reluplex (Katz et al., 2017) [60]
BaB (Bunel et al., 2018) [17]
Reluval (Wang et al., 2018) [118]
Neurify (Wang et al., 2018) [117]
AI² (Gehr et al., 2018) [45]
BaBSR (Bunel et al., 2019) [15]
MILP Based
Sherlock (Dutta et al., 2017) [31]
Reachability (Lomuscio and Maganti, 2017) [75]
ChengMIP (Cheng et al., 2017) [21]
Bounded MILP (Tjeng et al., 2019) [111]
Stability (Xiao et al., 2019) [127]
Others
DLV (Huang et al., 2017) [52]
GeoCert (Jordan et al., 2019) [58]
Linear Relaxation (§III-C)
Linear Programming
LP-Full (Weng et al., Salman et al., 2018, 2019) [121, 96]
C2V (Tjandraatmadja et al., 2020) [110]
Linear Inequality Propagation & Zonotopes
DiffAI (Mirman et al., 2018) [78]
Charon (Anderson et al., 2019) [1]
DeepZ (Singh et al., 2018) [101]
DeepPoly (Singh et al., 2019) [103]
Fast-Lin (Weng et al., 2018) [121]
CROWN (Zhang et al., 2018) [131]
CNN-Cert (Boopathy et al., 2019) [10]
Interval Bound Propagation
IBP (Gowal et al., 2019) [47]
AI² (Gehr et al., 2018) [45]
Linear Dual
WK,WK+ (Wong and Kolter, Wong et al., 2018, 2018) [122, 123]
PVT (Dvijotham et al., 2018) [33]
D-LP (Dvijotham et al., 2018) [34]
Hybrid Approaches
CROWN-IBP (Zhang et al., 2020) [134]
RefineZono (Singh et al., 2019) [104]
k-ReLU (Singh et al., 2019) [102]
Robust Training
Mixtrain (Wang et al., 2018) [116]
Robustra (Li et al., 2019) [72]
MMR (Croce et al., 2019) [26]
MMR-Universal (Croce and Hein, 2020) [25]
COLT (Balunovic and Vechev, 2020) [7]
SDP (§III-D)
SDP Train (Raghunathan et al., 2018) [90]
SDP Verify (Raghunathan et al., 2018) [91]
PGD-SDP (Dvijotham et al., 2019) [35]
LMIVerify (Fazlyab et al., 2019) [41]
Lipschitz & Curvature (§III-E)
CL (Hein and Andriushchenko, 2017) [51]
CRC/CRT (Singla and Feizi, 2020) [105]
Op-norm (Szegedy et al., 2013) [106]
Fast-Lip (Weng et al., 2018) [121]
RecurJac (Zhang et al., 2019) [133]
Differential Privacy
DP (Lecuyer et al., 2019) [65]
Rényi-divergence (Li et al., 2019) [71]
Neyman -Pearson Based
Rand. Smooth (Cohen et al., 2019) [24]
ℓ_0-Lee (Lee et al., 2019) [66]
ℓ_1-Teng (Teng et al., 2020) [108]
Zhang-Dual (Zhang et al., 2020) [130]
All Shapes (Yang et al., 2020) [128]
f-divergence (Dvijotham et al., 2020) [36]
Robust Training
Data Augmentation
Unlabeled (Carmon et al., 2019) [19]
Adversarial (Salman et al., 2019) [95]
Low Rank enforcement (Awasthi et al., 2020) [5]
Regularization
Consistency (Jeong and Shin, 2020) [54]
MACER (Zhai et al., 2020) [129]
ADRE (Feng et al., 2020) [43]

Certified Robustness for DNNs



https://github.com/AI-secure/Provable-Training-and-Verification-Approaches-Towards-Robust-Neural-Networks

Certified Robustness via Randomized Smoothing

- Neyman-Pearson lemma
- Smoothed classifier
- Certification bound
  - tightness

Related reading: Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification

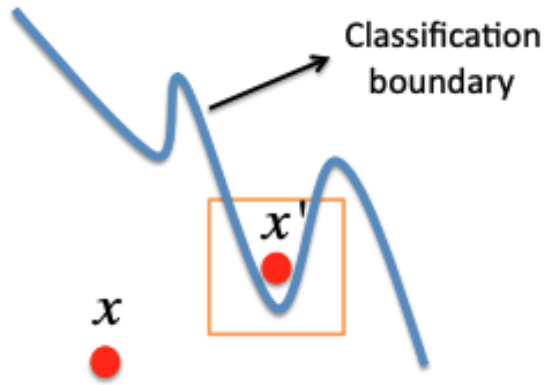


Illustration of the region-based classification. x is a testing benign example and x' is the corresponding adversarial example. The hypercube centered at x' intersects the most with the class region that has the true label.

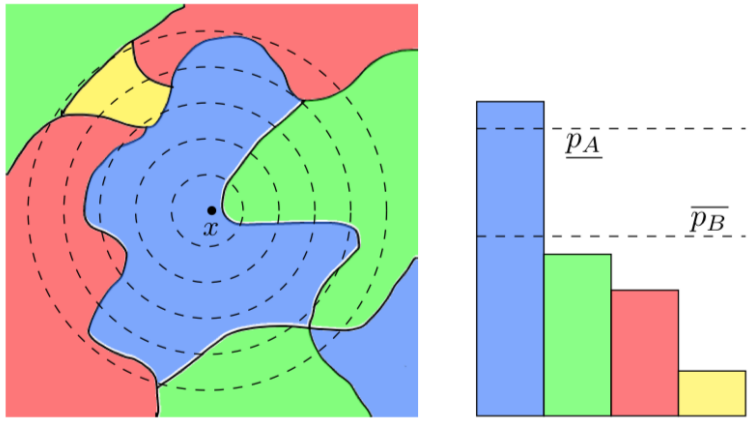
Algorithm 1 Learning Length r by Searching

Input: Validation dataset V, point-based DNN classifier C, step size ϵ, initial length r_0.

Output: Length r.

1: Initialize $r = r_0$.
2: ACC = Accuracy of C on V.
3: ACC_{RC} = Accuracy of the $RC_{C,r}$ classifier on V.
4: **while** $ACC_{RC} \geq ACC$ **do**
5: $r = r + \epsilon$.
6: ACC_{RC} = Accuracy of the $RC_{C,r}$ classifier on V.
7: **end while**
8: **return** $r - \epsilon$.

Certified Robustness via Randomized Smoothing



Smoothed classifier:

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\varepsilon) = c)$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

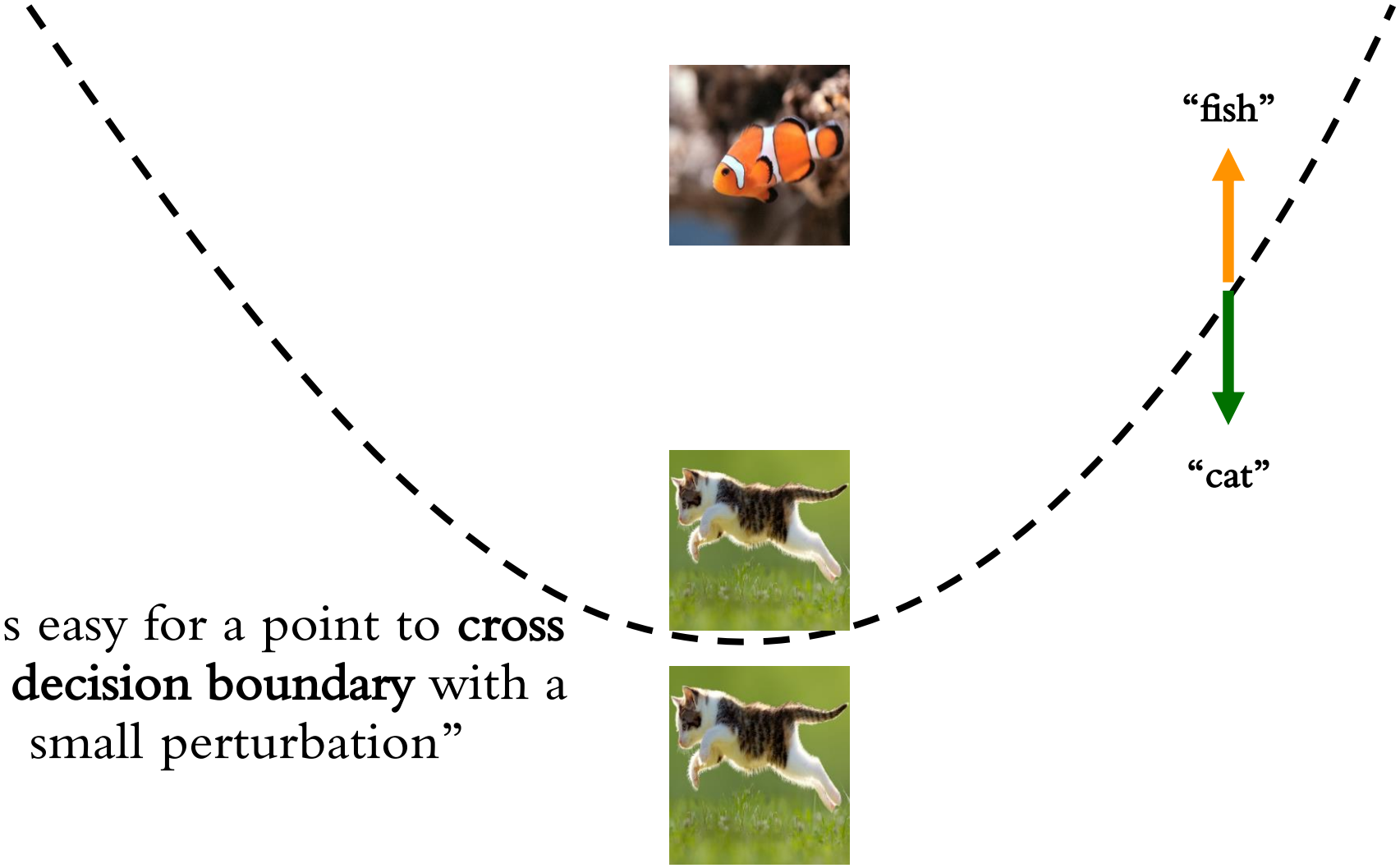
Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c)$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

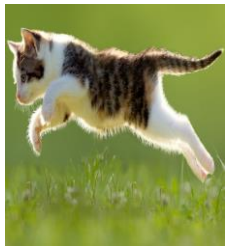
$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$$

Adversarial Attacks: Decision Boundary Intuition



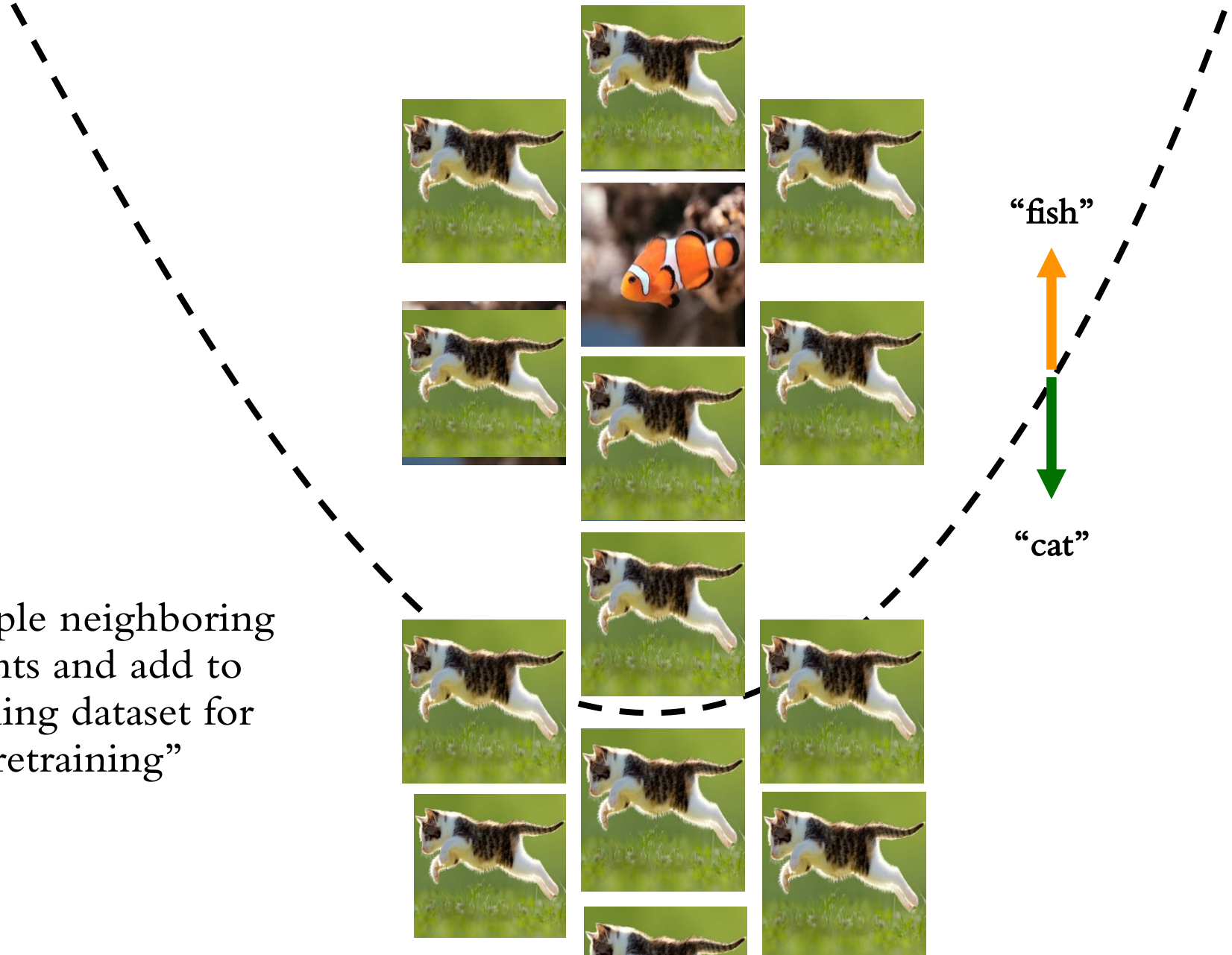
"It's easy for a point to **cross the decision boundary** with a small perturbation"

Defense: Adversarial Retraining



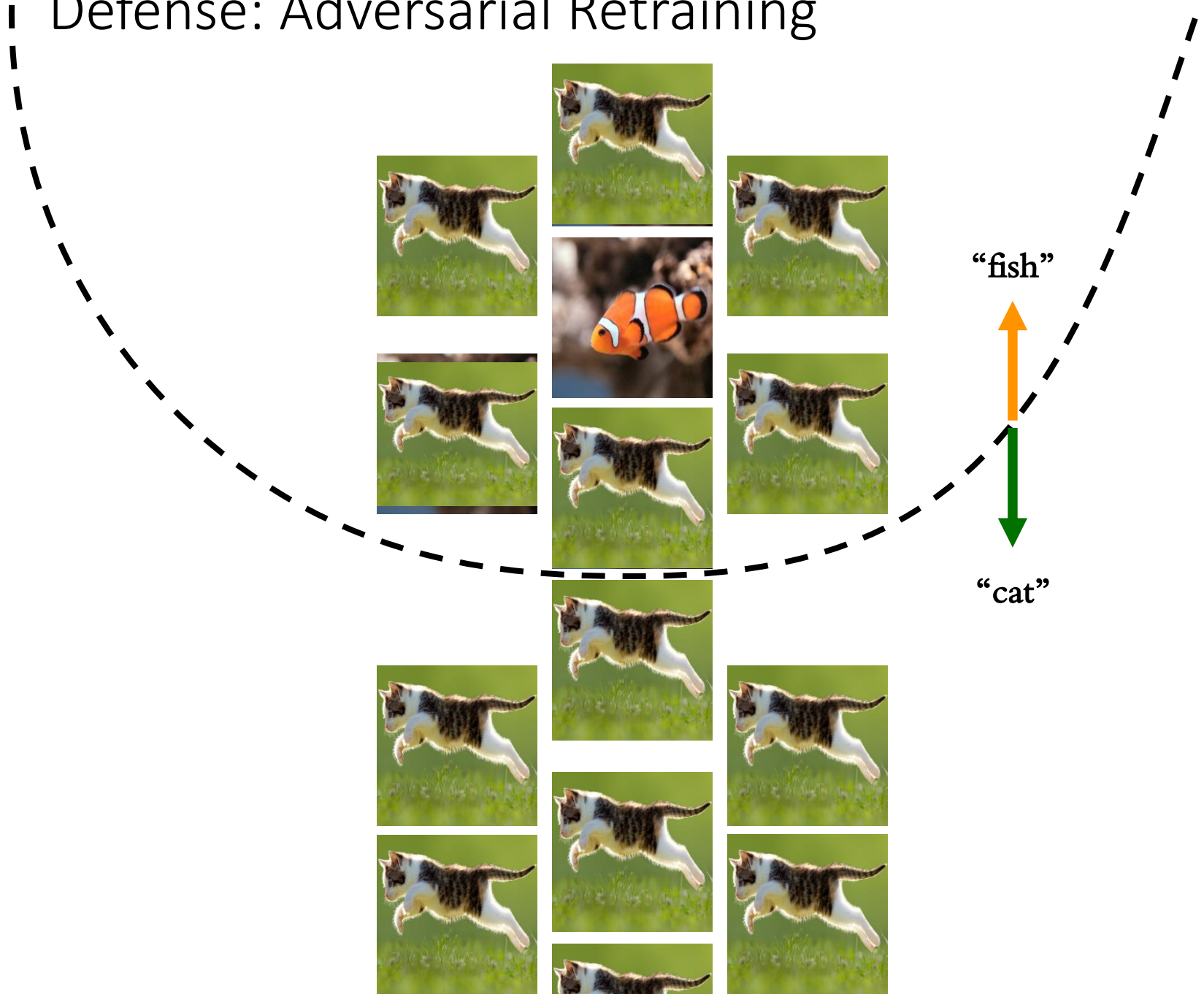
Robustness: "The points **near** (by some distance measure) a point should have the **same label**"

Defense: Adversarial Retraining

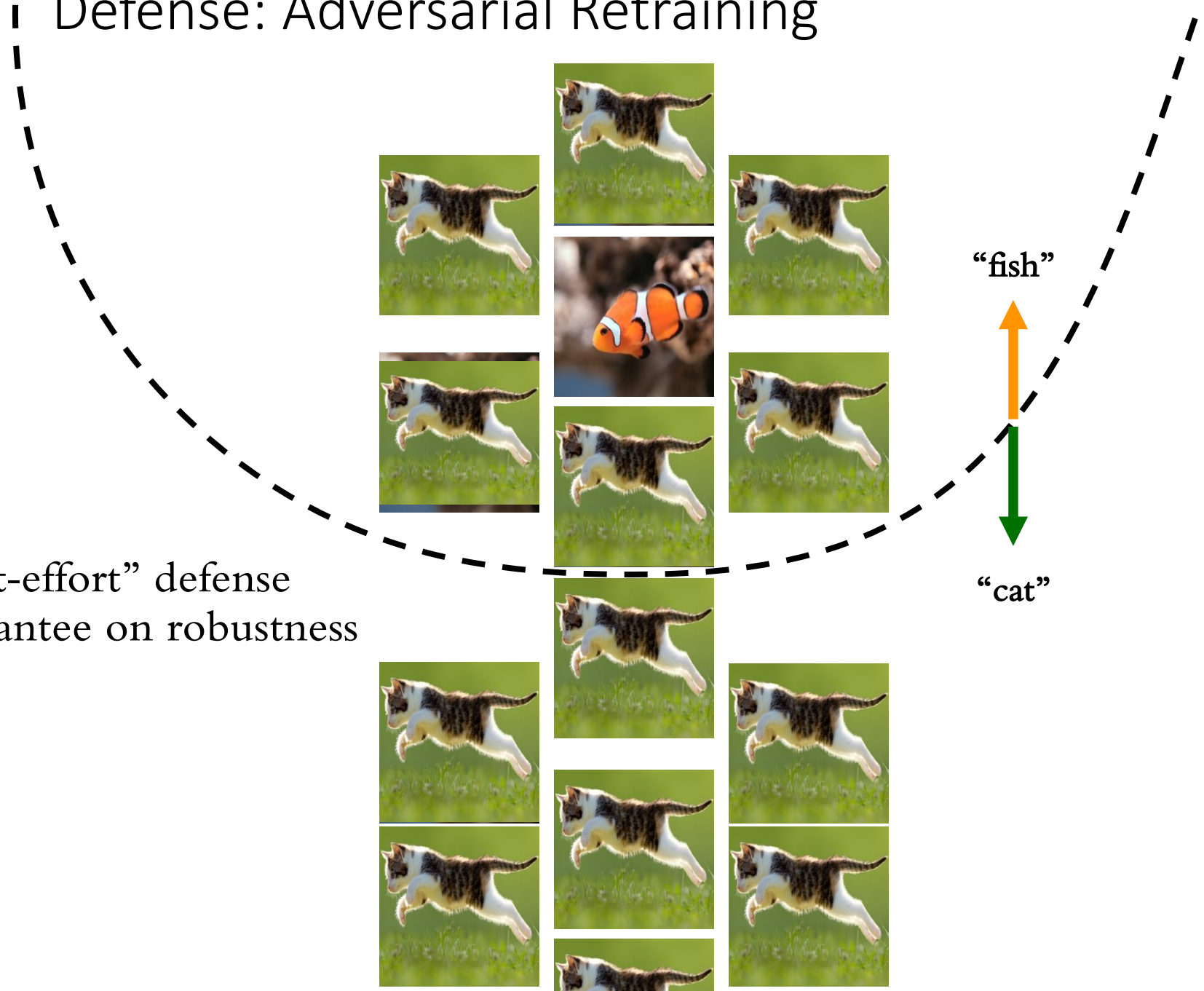


"Sample neighboring points and add to training dataset for retraining"

Defense: Adversarial Retraining

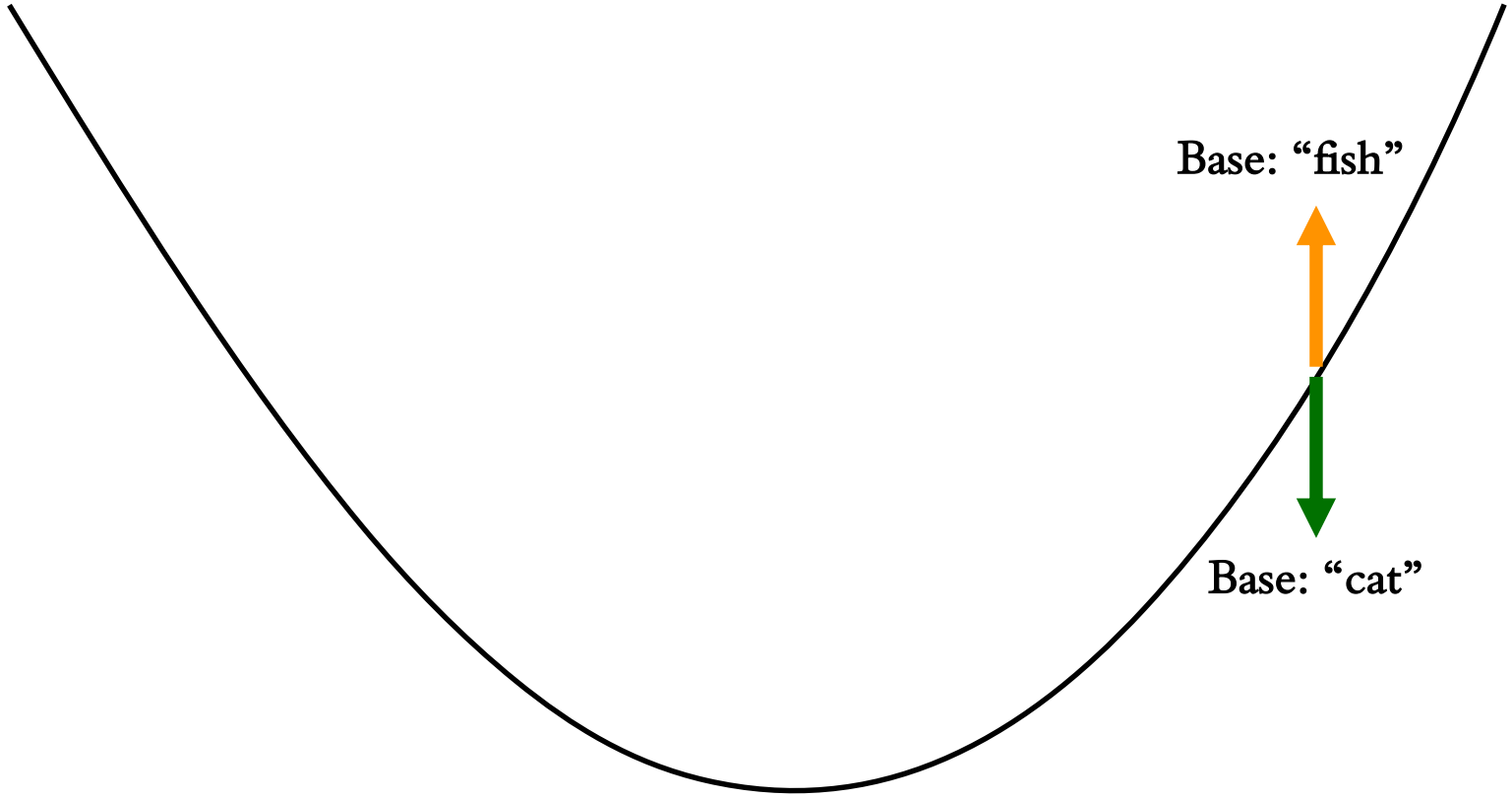


Defense: Adversarial Retraining



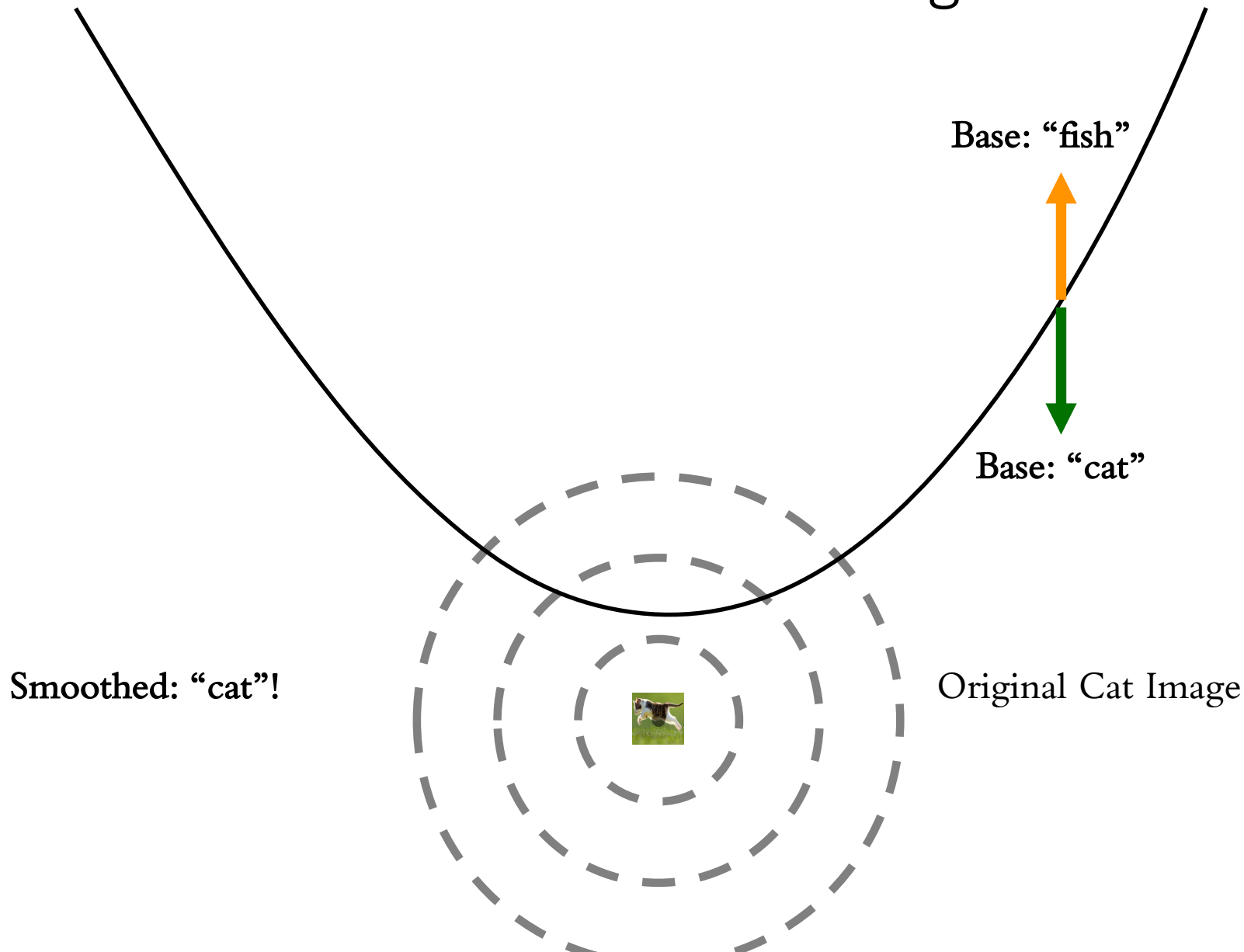
"Best-effort" defense
No guarantee on robustness

Defense: Randomized Smoothing

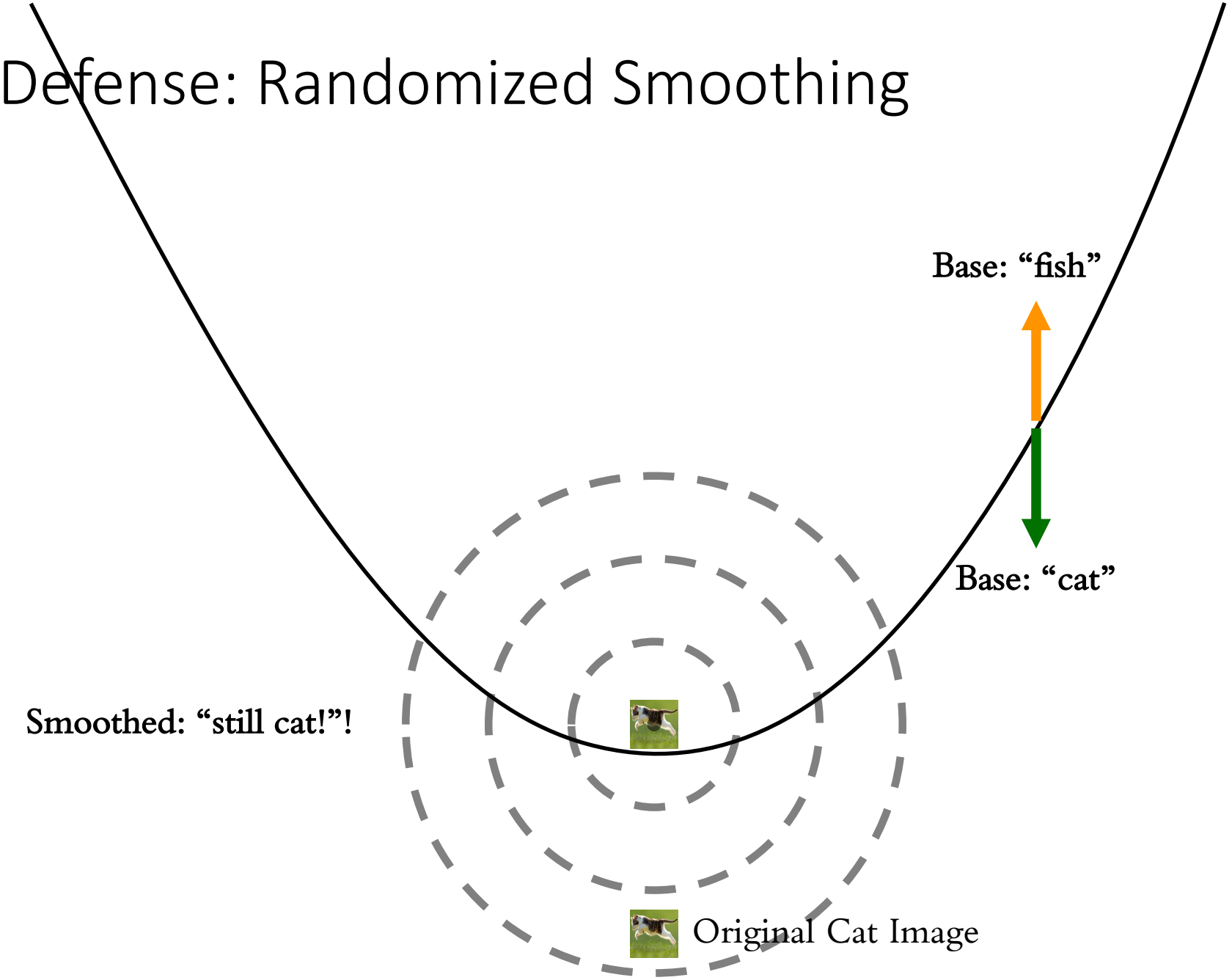


 Original Cat Image

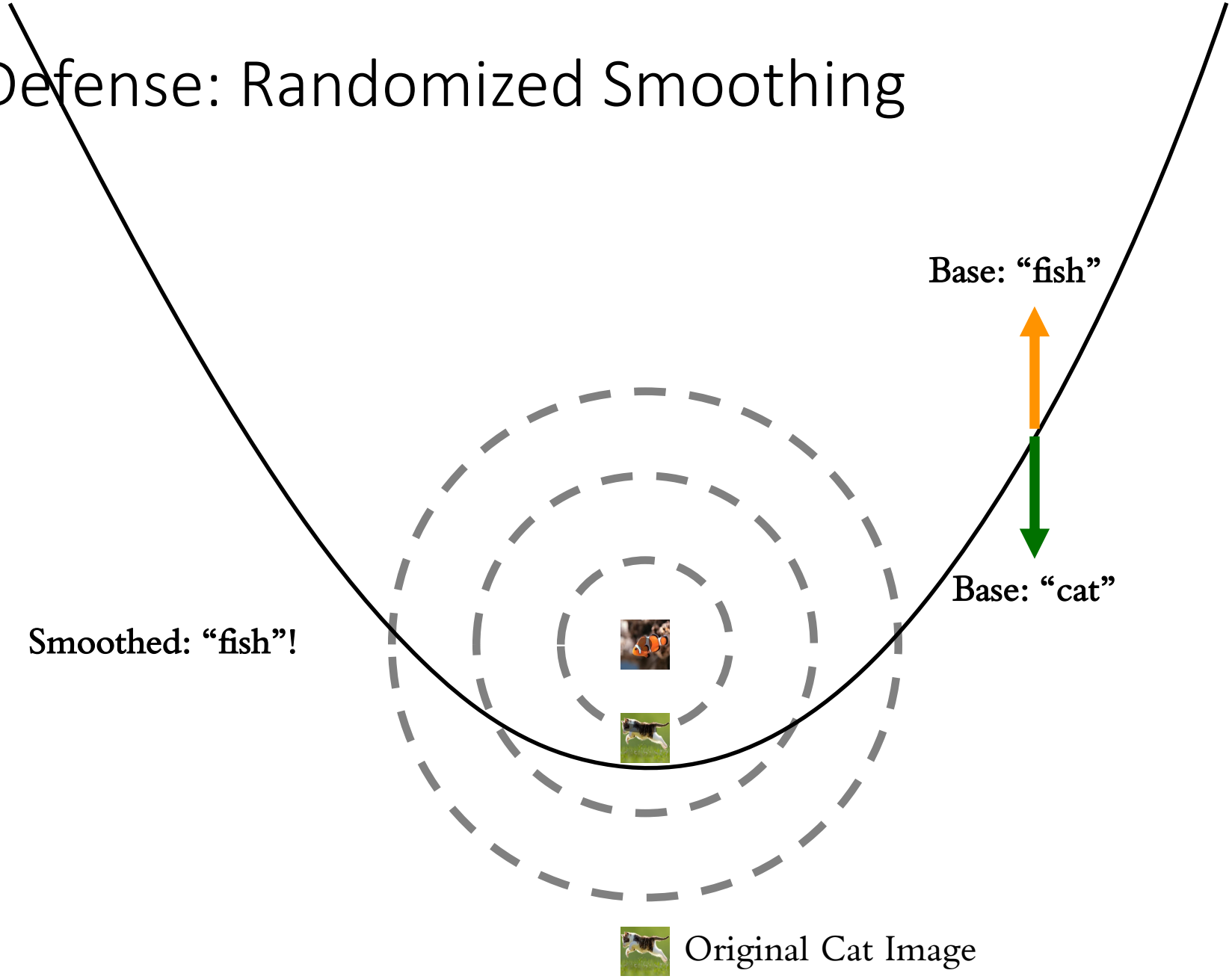
Defense: Randomized Smoothing



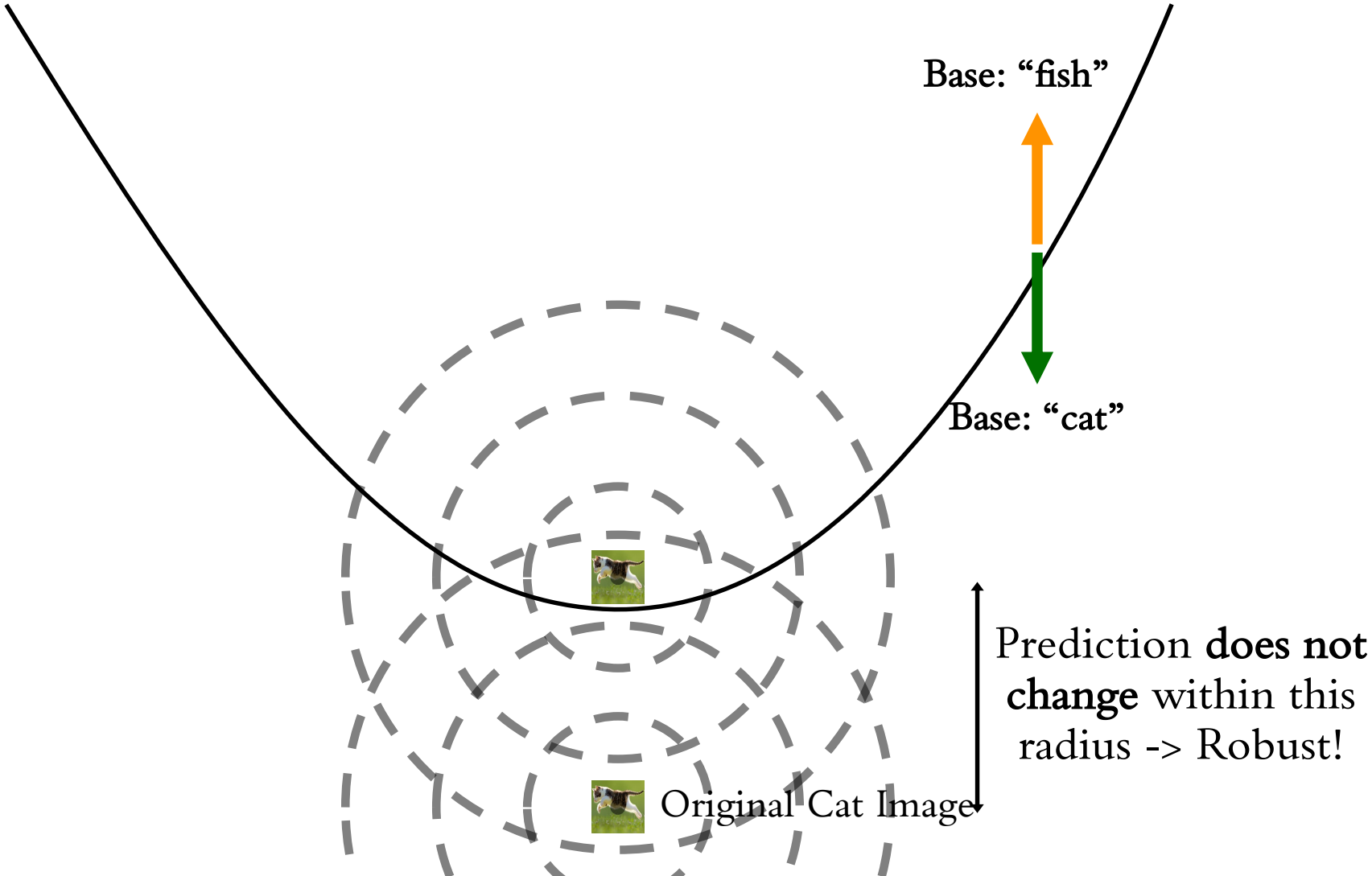
Defense: Randomized Smoothing



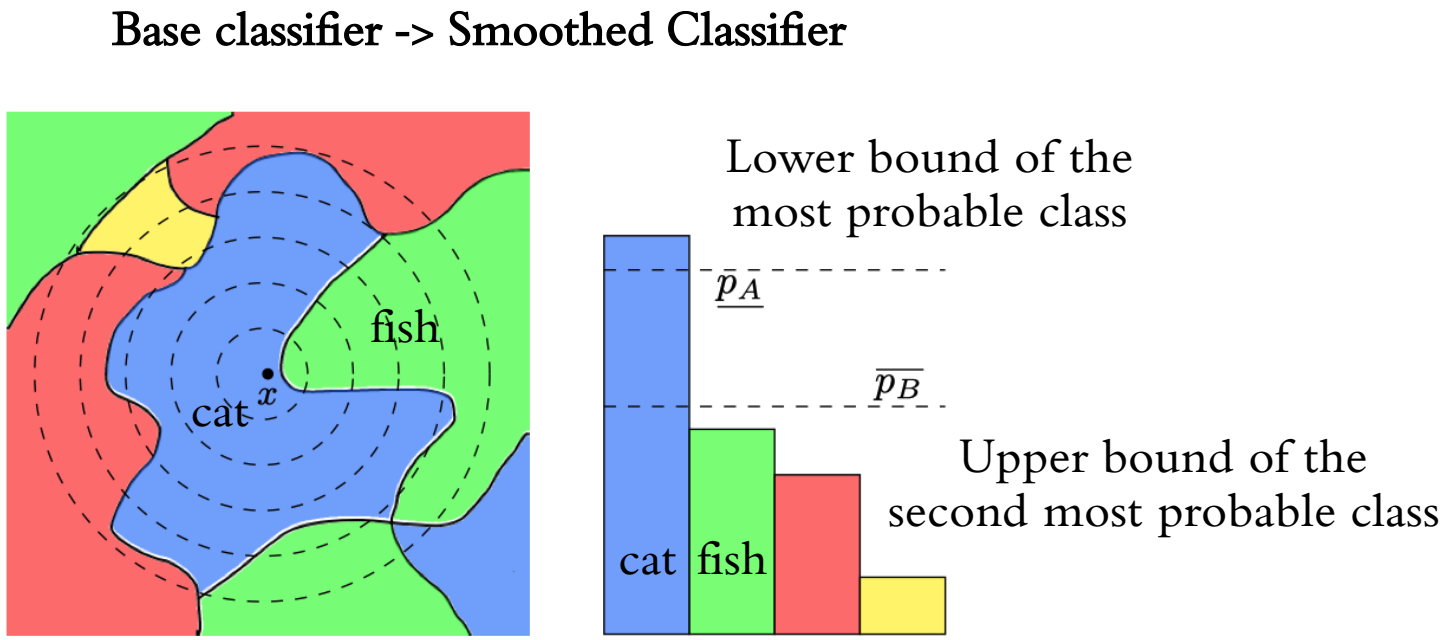
Defense: Randomized Smoothing



Defense: Randomized Smoothing



High-level Intuition of Randomized Smoothing



- Goal: "Smooth" out the classifier, use the class that takes the largest proportion of the predictions (by base classifier) in the Gaussian ball around the given point x as prediction of smoothed classifier.

Image Credit: Certified Adversarial Robustness via Randomized Smoothing

Robustness Guarantee

Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy:*

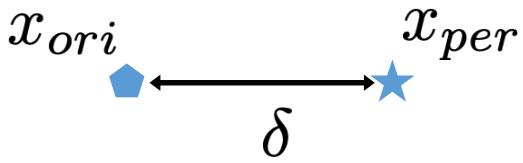
$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (2)$$

Then $g(x + \delta) = c_A$ for all $\|\delta\|_2 < R$, where

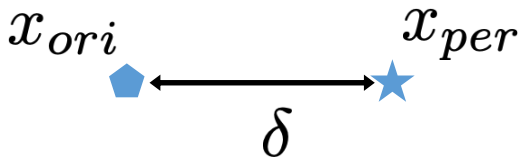
$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (3)$$

- No assumption on f
- Certified radius R is large when: noise level is high; pA is large; pB is small. When pA is close to 1, R goes to infinity.

(Informal) Understanding of Robustness Guarantee

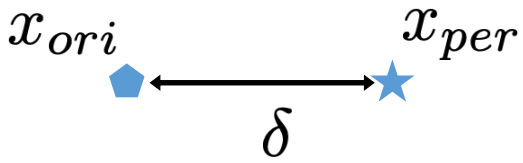


(Informal) Understanding of Robustness Guarantee



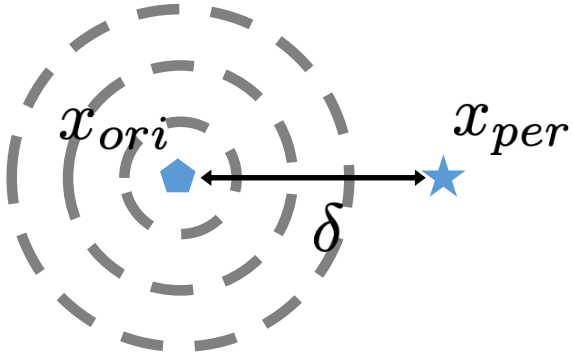
Two points: the original image and the perturbed image

(Informal) Understanding of Robustness Guarantee



Robust: We would like these two points to have the **same label** under the prediction of **smoothed classifier**

(Informal) Understanding of Robustness Guarantee



Probability of x_{ori} predicted as class A

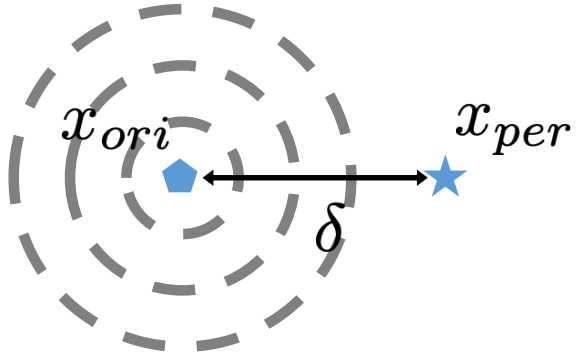
$$p_{c_A}(x_{ori}) = \int_{x \sim x_{ori} + \mathcal{N}(0,\sigma^2 I)} I_{c_A}(f(x)) \mu_{ori}(x) dx$$

Probability Density Function

Indicator Function

I = 1 if arg max f(x) is index of class A

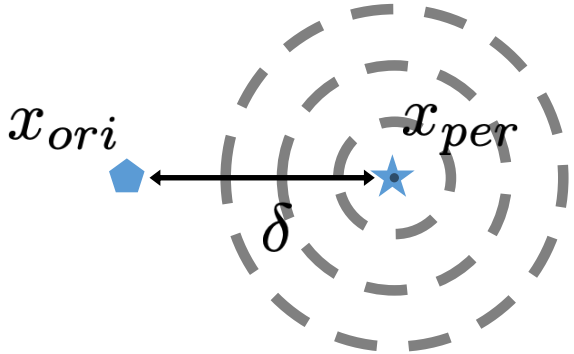
(Informal) Understanding of Robustness Guarantee



$$p_{c_A}(x_{ori}) = \int_{x \sim x_{ori} + \mathcal{N}(0, \sigma^2 I)} I_{c_A}(f(x)) \mu_{ori}(x) dx$$

$$= \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0$$

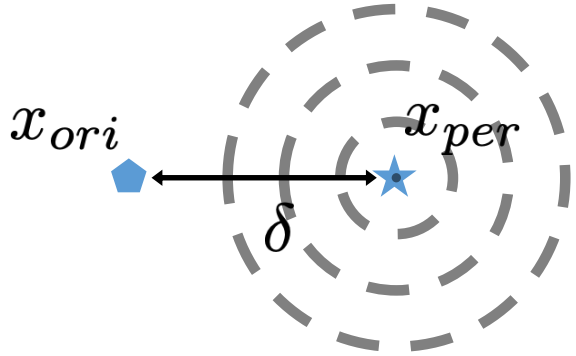
(Informal) Understanding of Robustness Guarantee



Probability of x_{per} predicted as class A

$$p_{c_A}(x_{per}) = \int_{x \sim x_{per} + \mathcal{N}(0,\sigma^2 I)} I_{c_A}(f(x)) \mu_{per}(x) dx$$

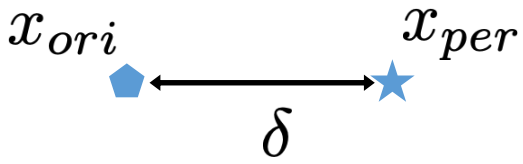
(Informal) Understanding of Robustness Guarantee



$$p_{c_A}(x_{per}) = \int_{x \sim x_{per} + \mathcal{N}(0, \sigma^2 I)} I_{c_A}(f(x)) \mu_{per}(x) dx$$

$$= \int_{x_\delta \sim \mathcal{N}(x_{ori} + \delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta$$

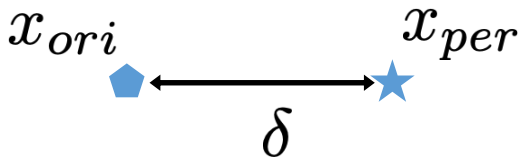
(Informal) Understanding of Robustness Guarantee



$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta$$

(Informal) Understanding of Robustness Guarantee

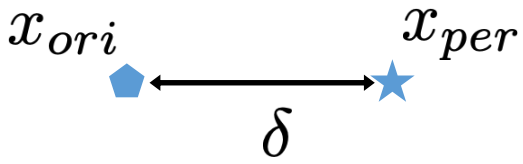


$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta \quad ?$$

Seek the **worst case scenario**!

(Informal) Understanding of Robustness Guarantee

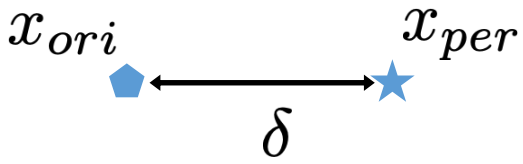


$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta \quad (*)$$

worst case scenario:
"the lower bound of (*)"

(Informal) Understanding of Robustness Guarantee



$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta \quad (*)$$

Worst case scenario Intuition:

for a given x, when $I(f(x)) = 1$,
$\mu_{ori}(x)$ is large and $\mu_{per}(x)$ is small,
vice versa

-> Neyman Pearson Lemma!

(Intuitive) Connection between What We Want and Neyman-Pearson

$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta \quad (*)$$

For a given x, when $I(f(x)) = 1, \mu_{ori}(x)$ is large and $\mu_{per}(x)$ is small, vice versa

Neyman-Pearson: **Bound ratio between the two densities**

(Intuitive) Connection between What We Want and Neyman-Pearson

$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta \quad (*)$$

For a given x, when $I(f(x)) = 1, \mu_{ori}(x)$ is large and $\mu_{per}(x)$ is small, vice versa

Neyman-Pearson: **Bound ratio between the two densities**

def $S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$

if $\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$

When the original image is predicted correctly with high probability,

then $\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$

The perturbed image will also be predicted correctly with high probability.

Backup Slide

Neyman-Pearson Lemma Proof

$$S = \left\{ z \in \mathbb{R}^d : \frac{\mu_Y(z)}{\mu_X(z)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(h(X) = 1) \geq \mathbb{P}(X \in S)$$

$$\mathbb{P}(h(Y) = 1) \geq \mathbb{P}(Y \in S)$$

$$\begin{aligned}
&\mathbb{P}(h(Y) = 1) - \mathbb{P}(Y \in S) \\
=&\int_{\mathbb{R}^d} h(1|z)\mu_Y(z)dz - \int_S \mu_Y(z)dz \\
=&\left[\int_{S^c} h(1|z)\mu_Y(z)dz + \int_S h(1|z)\mu_Y(z)dz\right] - \left[\int_S h(1|z)\mu_Y(z)dz + \int_S h(0|z)\mu_Y(z)dz\right] \\
=&\int_{S^c} h(1|z)\mu_Y(z)dz - \int_S h(0|z)\mu_Y(z)dz \\
\geq& t\left[\int_{S^c} h(1|z)\mu_X(z)dz - \int_S h(0|z)\mu_X(z)dz\right] \\
=& t\left[\int_{S^c} h(1|z)\mu_X(z)dz + \int_S h(1|z)\mu_X(z)dz - \int_S h(1|z)\mu_X(z)dz - \int_S h(0|z)\mu_X(z)dz\right] \\
=& t\left[\int_{\mathbb{R}^d} h(1|z)\mu_X(z)dz - \int_S \mu_X(z)dz\right] \\
=& t\left[\mathbb{P}(h(X) = 1) - \mathbb{P}(X \in S)\right] \\
\geq& 0
\end{aligned}$$

Use Neyman-Pearson to show Robustness Condition

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

$$\begin{aligned}
\frac{\mu_{per}(x)}{\mu_{ori}(x)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-(\delta+x_{ori}))^2}{2\sigma^2}\right)}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-x_{ori})^2}{2\sigma^2}\right)} \\
&= \exp\left(\frac{1}{\sigma^2}\delta^T x - \frac{2\delta^T x_{ori} + \|\delta\|^2}{2\sigma^2}\right) \\
&= \exp\left(a\delta^T x + b\right) \\
a &= \frac{1}{\sigma^2} \\
b &= -\frac{2\delta^T x_{ori} + \|\delta\|^2}{2\sigma^2}
\end{aligned}$$

$$t = \exp(a\beta + b)$$

$$\frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \Leftrightarrow \exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

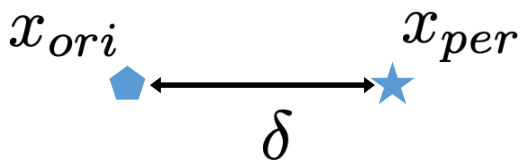
(Revisit) Robustness Theorem

Theorem 1 (restated). *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function. Let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let $g(x) = \arg\max_c \mathbb{P}(f(x+\varepsilon) = c)$. Suppose that for a specific $x \in \mathbb{R}^d$, there exist $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ such that:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \tag{6}$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \tag{7}$$



$$x_0 \in X := x_{ori} + \varepsilon = \mathcal{N}(x_{ori}, \sigma^2 I)$$
$$x_\delta \in Y := x_{ori} + \delta + \varepsilon = \mathcal{N}(x_{ori} + \delta, \sigma^2 I)$$

Known: $\mathbb{P}(f(X) = c_A) \geq \underline{p_A}$ and $\mathbb{P}(f(X) = c_B) \leq \overline{p_B}$

Need to show: $\mathbb{P}(f(Y) = c_A) > \mathbb{P}(f(Y) = c_B)$

(Formal) Proof of Robustness Guarantee

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

Condition of set S:

$$t = \exp(a\beta + b)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

(Formal) Proof of Robustness Guarantee

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

Condition of set S:

$$t = \exp(a\beta + b)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

Construct: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

Easy to show: $\mathbb{P}(X \in A) = \underline{p_A}$

$$\begin{aligned}
\mathbb{P}(X \in A) &= \mathbb{P}(\delta^T (X - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(\delta^T \mathcal{N}(0, \sigma^2 I) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(\|\delta\|(\sigma Z) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(Z \leq \Phi^{-1}(\underline{p_A}))
\end{aligned}$$

(Formal) Proof of Robustness Guarantee

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

$$t = \exp(a\beta + b)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

Define: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

Easy to show: $\mathbb{P}(X \in A) = \underline{p_A}$

By definition:

$$\mathbb{P}(f(X) = c_A) \geq \underline{p_A}$$

(Formal) Proof of Robustness Guarantee

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$t = \exp(a\beta + b)$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

Define: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

Easy to show: $\mathbb{P}(X \in A) = \underline{p_A}$

Given:

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) = \mathbb{P}(f(X) = c_A) \geq \underline{p_A} = \mathbb{P}(X \in A) = \mathbb{P}(x_0 \in S)$$

(Proof) Lower Bound of Correct Class A

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$t = \exp(a\beta + b)$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

Define: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

$$\mathbb{P}(X \in A) = \underline{p_A}$$

Given:

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) = \mathbb{P}(f(X) = c_A) \geq \underline{p_A} = \mathbb{P}(X \in A) = \mathbb{P}(x_0 \in S)$$

By Neyman-Pearson Get:

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

(Proof) Lower Bound of Correct Class A

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$t = \exp(a\beta + b)$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

Define: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

$$\mathbb{P}(X \in A) = \underline{p_A}$$

Given:

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) = \mathbb{P}(f(X) = c_A) \geq \underline{p_A} = \mathbb{P}(X \in A) = \mathbb{P}(x_0 \in S)$$

By Neyman-Pearson + definition:

$$\mathbb{P}(f(Y) = c_A) = \mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S) = \mathbb{P}(Y \in A)$$

(Proof) Lower Bound of Correct Class A

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \leq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) \geq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S)$$

$$t = \exp(a\beta + b)$$

$$\exp\left(a\delta^T x + b\right) \leq t \Leftrightarrow \delta^T x \leq \beta$$

Define: $A := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(\underline{p_A}) \right\}$

$$\mathbb{P}(X \in A) = \underline{p_A}$$

Given:

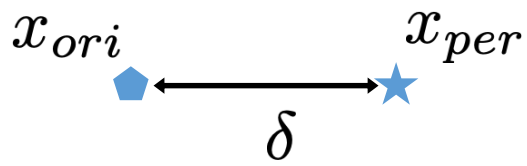
$$\mathbb{P}(I_{c_A}(f(x_0)) = 1) = \mathbb{P}(f(X) = c_A) \geq \underline{p_A} = \mathbb{P}(X \in A) = \mathbb{P}(x_0 \in S)$$

By Neyman-Pearson + definition:

$$\mathbb{P}(f(Y) = c_A) = \mathbb{P}(I_{c_A}(f(x_\delta)) = 1) \geq \mathbb{P}(x_\delta \in S) = \mathbb{P}(Y \in A)$$

We get the **lower bound**!

(Recall) The Other Direction



Correct Class A

$$p_{ori} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_A}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_A$$

$$p_{per} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_A}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta$$

Other Classes B

$$p_{ori,B} = \int_{x_0 \sim \mathcal{N}(x_{ori}, \sigma^2 I)} I_{c_B}(f(x_0)) \mu_{ori}(x_0) dx_0 \quad = p_B$$

$$p_{per,B} = \int_{x_\delta \sim \mathcal{N}(x_{ori}+\delta, \sigma^2 I)} I_{c_B}(f(x_\delta)) \mu_{per}(x_\delta) dx_\delta$$

(Proof) Upper Bound of Other Class B

$$S = \left\{ x \in \mathbb{R}^d : \frac{\mu_{per}(x)}{\mu_{ori}(x)} \geq t \right\}, \exists t > 0$$

$$\mathbb{P}(I_{c_B}(f(x_0)) = 1) \leq \mathbb{P}(x_0 \in S)$$

$$\mathbb{P}(I_{c_B}(f(x_\delta)) = 1) \leq \mathbb{P}(x_\delta \in S)$$

$$t = \exp(a\beta + b)$$

$$\exp\left(a\delta^T x + b\right) \geq t \Leftrightarrow \delta^T x \geq \beta$$

Define: $B := \left\{ x : \delta^T (x - x_{ori}) \leq \sigma \|\delta\| \Phi^{-1}(1 - \overline{p_B}) \right\}$

$$\mathbb{P}(X \in B) = \overline{p_B}$$

Given:

$$\mathbb{P}(I_{c_B}(f(x_\delta)) = 1) = \mathbb{P}(f(Y) = c_B) \leq \overline{p_B} = \mathbb{P}(Y \in B) = \mathbb{P}(x_\delta \in S)$$

Get:

$$\mathbb{P}(f(Y) = c_B) = \mathbb{P}(I_{c_B}(f(x_\delta)) = 1) \leq \mathbb{P}(x_\delta \in S) = \mathbb{P}(Y \in B)$$

Almost repeat the proof -> upper bound

(Formal) Proof of Robustness Guarantee

We have:

$$\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A)$$

$$\mathbb{P}(Y \in B) \geq \mathbb{P}(f(Y) = c_B)$$

We want:

$$\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A) > \mathbb{P}(Y \in B) \geq \mathbb{P}(f(Y) = c_B)$$

This is the missing component!

(Formal) Proof of Robustness Guarantee

$$
\begin{aligned}
\mathbb{P}(Y \in A) &= \mathbb{P}(\delta^T(Y - x_{ori}) \leq \sigma\|\delta\|\Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(\delta^T\mathcal{N}(\delta, \sigma^2 I) \leq \sigma\|\delta\|\Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(\|\delta\|(\sigma Z + \|\delta\|) \leq \sigma\|\delta\|\Phi^{-1}(\underline{p_A})) \\
&= \mathbb{P}(Z \leq \Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|}{\sigma}) \\
&= \Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|}{\sigma}\right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(Y \in B) &= \mathbb{P}(\delta^T(Y - x_{ori}) \geq \sigma\|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&= \mathbb{P}(\delta^T\mathcal{N}(\delta, \sigma^2 I) \geq \sigma\|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&= \mathbb{P}(\|\delta\|(\sigma Z + \|\delta\|) \geq \sigma\|\delta\|\Phi^{-1}(1 - \overline{p_B})) \\
&= \mathbb{P}(Z \geq \Phi^{-1}(1 - \overline{p_B}) - \frac{\|\delta\|}{\sigma}) \\
&= \mathbb{P}(Z \leq \Phi^{-1}(\overline{p_B}) + \frac{\|\delta\|}{\sigma}) \\
&= \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\delta\|}{\sigma}\right)
\end{aligned}
$$

(Formal) Proof of Robustness Guarantee

We have:

$$\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A)$$

$$\mathbb{P}(Y \in B) \geq \mathbb{P}(f(Y) = c_B)$$

We want:

$$\mathbb{P}(f(Y) = c_A) \geq \mathbb{P}(Y \in A) > \mathbb{P}(Y \in B) \geq \mathbb{P}(f(Y) = c_B)$$

$$\mathbb{P}(Y \in A) = \Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|}{\sigma}\right) > \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\delta\|}{\sigma}\right) = \mathbb{P}(Y \in B)$$

$$\|\delta\| < \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right)$$

∎ Proof Done.

Experiments: Training

- Method: train the base classifier with Gaussian data augmentation at variance σ^2

The log-probabilities that f classifies each noisy point as the ground truth label of the clean point

$$\sum_{i=1}^{n} \log \mathbb{P}_\varepsilon(f(x_i+\varepsilon)=c_i) = \sum_{i=1}^{n} \log \mathbb{E}_\varepsilon 1[\arg\max_c f_c(x_i+\varepsilon)=c_i]$$

$$\approx \sum_{i=1}^{n} \log \mathbb{E}_\varepsilon [\frac{\exp(f_{c_i}(x_i+\varepsilon))}{\sum_{c\in\mathcal{Y}}\exp(f_c(x_i+\varepsilon))}]$$

$$\geq \sum_{i=1}^{n} \mathbb{E}_\varepsilon [\log \frac{\exp(f_{c_i}(x_i+\varepsilon))}{\sum_{c\in\mathcal{Y}}\exp(f_c(x_i+\varepsilon))}]$$

Experiments: Training

- Method: train the base classifier with Gaussian data augmentation at variance σ^2

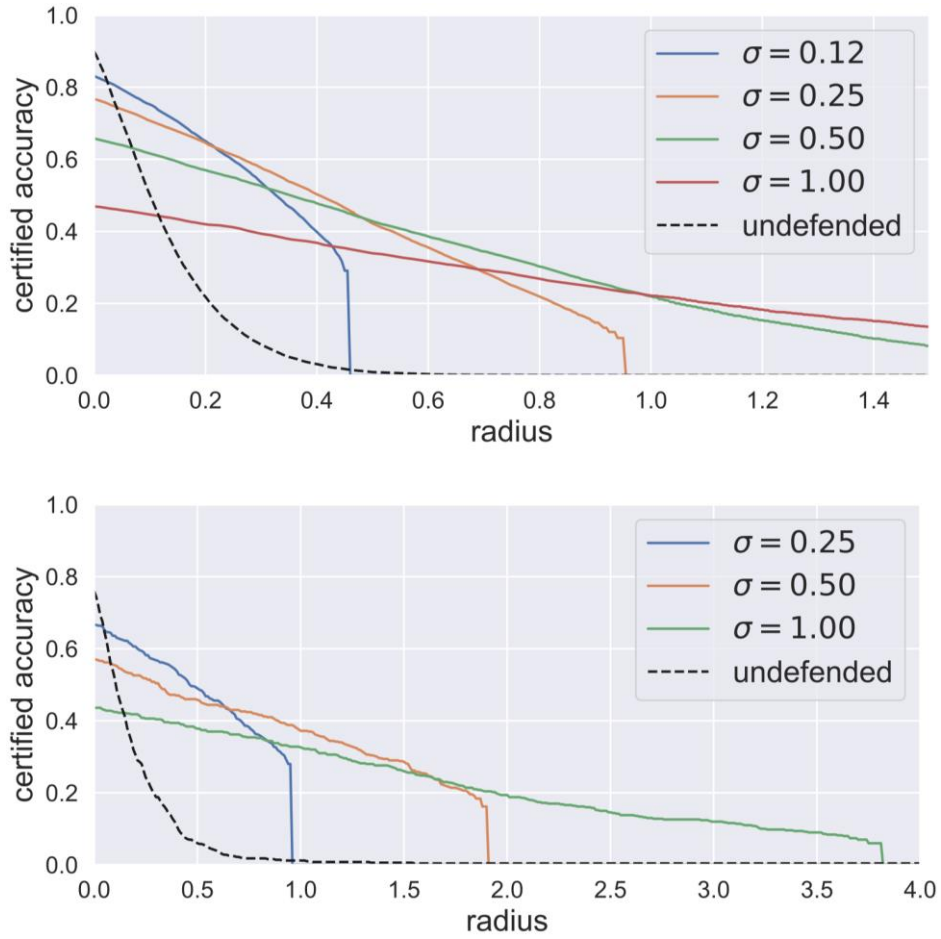
$$\sum_{i=1}^{n} \log \mathbb{P}_\varepsilon(f(x_i+\varepsilon)=c_i) = \sum_{i=1}^{n} \log \mathbb{E}_\varepsilon 1[\arg\max_c f_c(x_i+\varepsilon)=c_i]$$

$$\approx \sum_{i=1}^{n} \log \mathbb{E}_\varepsilon [\frac{\exp(f_{c_i}(x_i+\varepsilon))}{\sum_{c\in\mathcal{Y}} \exp(f_c(x_i+\varepsilon))}]$$

[Jensen's inequality & concavity of log] $$\geq \sum_{i=1}^{n} \mathbb{E}_\varepsilon [\log \frac{\exp(f_{c_i}(x_i+\varepsilon))}{\sum_{c\in\mathcal{Y}} \exp(f_c(x_i+\varepsilon))}]$$

Negative of the cross-entropy loss under Gaussian data augmentation

Approximate Certified Accuracy



- Robustness/accuracy tradeoff
  - σ low -> small radii certified with high accuracy, but large radii cannot be certified.
  - σ high -> larger radii can be certified, but smaller radii are certified at a lower accuracy.

Approximate certified accuracy attained by randomized smoothing on CIFAR-10 (top) and ImageNet (bottom)

Image Credit: Certified Adversarial Robustness via Randomized Smoothing

Intuition: Linear Classifier as Worst Case

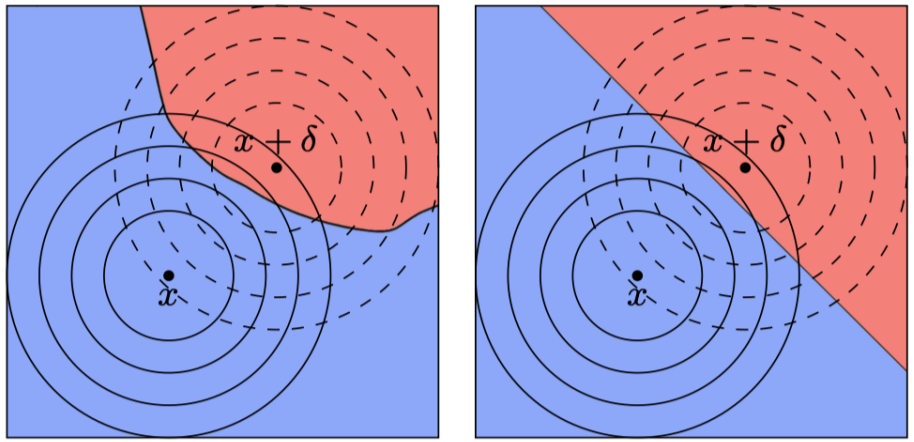
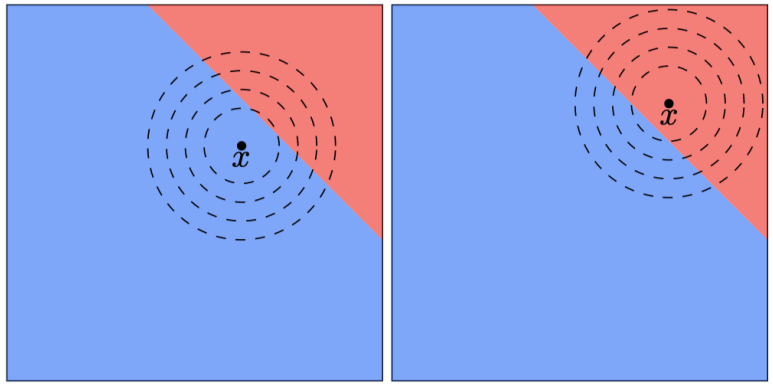


Figure 3. Illustration of f^* in two dimensions. The concentric circles are the density contours of $\mathcal{N}(x, \sigma^2 I)$ and $\mathcal{N}(x + \delta, \sigma^2 I)$. Out of all base classifiers f which classify $\mathcal{N}(x, \sigma^2 I)$ as c_A (blue) with probability $\geq \underline{p_A}$, such as both classifiers depicted above, the "worst-case" f^* — the one which classifies $\mathcal{N}(x + \delta, \sigma^2 I)$ as c_A with minimal probability — is depicted on the right: a linear classifier with decision boundary normal to the perturbation δ.

"The certified bound for a linear two-class classifier is tight"

Image Credit: Certified Adversarial Robustness via Randomized Smoothing

Smoothing a Two-Class Linear Classifier



$$f(x) = \text{sign}(w^T x + b)$$

Intuition: An isotropic Gaussian will put more mass on whichever half-space its center x lies in. So smoothing does not change decision for any point.

Proposition 3. *If f is a two-class linear classifier $f(x) = sign(w^T x + b)$, and g is the smoothed version of f with any σ, then $g(x) = f(x)$ for any x (where f is defined).*

$$\begin{aligned}
g(x) = 1 &\iff \mathbb{P}_\varepsilon(f(x+\varepsilon) = 1) > \frac{1}{2} && (\varepsilon \sim \mathcal{N}(0, \sigma^2 I)) \\
&\iff \mathbb{P}_\varepsilon\left(\text{sign}(w^T(x+\varepsilon) + b) = 1\right) > \frac{1}{2} \\
&\iff \mathbb{P}_\varepsilon\left(w^T x + w^T\varepsilon + b \geq 0\right) > \frac{1}{2} \\
&\iff \mathbb{P}\left(\sigma\|w\| Z \geq -w^T x - b\right) > \frac{1}{2} && (Z \sim \mathcal{N}(0,1)) \\
&\iff \mathbb{P}\left(Z \leq \frac{w^T x + b}{\sigma\|w\|}\right) > \frac{1}{2} \\
&\iff \frac{w^T x + b}{\sigma\|w\|} > 0 \\
&\iff w^T x + b > 0 \\
&\iff f(x) = 1
\end{aligned}$$

the other direction (-1) is similar

Two-Class Linear Classifier Certified Radius

Proposition 4. *If f is a two-class linear classifier $f(x) = sign(w^T x + b)$, and g is the smoothed version of f with any σ, then invoking Theorem 1 at any x (where f is defined) with $\underline{p_A} = p_A$ and $\overline{p_B} = p_B$ will yield the certified radius $R = \frac{|w^T x+b|}{\|w\|}$.*

Thm1: $R = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right)$

Proof. In binary classification, $p_A = 1 - p_B$, so Theorem 1 returns $R = \sigma\Phi^{-1}(\underline{p_A})$.

We have:

$$
\begin{aligned}
p_A &= \mathbb{P}_\varepsilon(f(x+\varepsilon) = g(x)) \\
&= \mathbb{P}_\varepsilon(\text{sign}(w^T(x+\varepsilon)+b) = \text{sign}(w^T x + b)) \qquad \text{Prop3: } g(x) = f(x) \\
&= \mathbb{P}_\varepsilon(\text{sign}(w^T x + \sigma\|w\|Z + b) = \text{sign}(w^T x + b))
\end{aligned}
$$

$w^T x + b > 0$

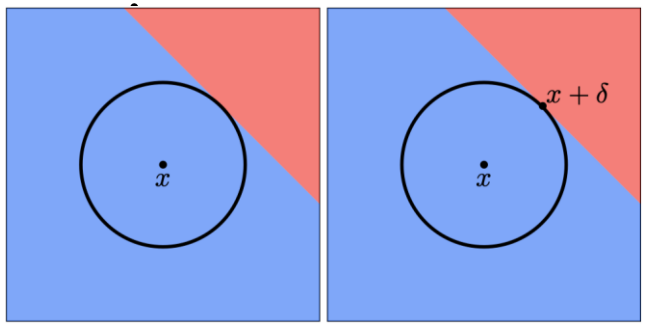
$$
\begin{aligned}
p_A &= \mathbb{P}_\varepsilon(w^T x + \sigma\|w\|Z + b > 0) \\
&= \mathbb{P}_\varepsilon\left(Z > \frac{-w^T x - b}{\sigma\|w\|}\right) \\
&= \mathbb{P}_\varepsilon\left(Z < \frac{w^T x + b}{\sigma\|w\|}\right) \\
&= \Phi\left(\frac{w^T x + b}{\sigma\|w\|}\right)
\end{aligned}
$$

$w^T x + b < 0$

$$
\begin{aligned}
p_A &= \mathbb{P}_\varepsilon(w^T x + \sigma\|w\|Z + b < 0) \\
&= \mathbb{P}_\varepsilon\left(Z < \frac{-w^T x - b}{\sigma\|w\|}\right) \\
&= \Phi\left(\frac{-w^T x - b}{\sigma\|w\|}\right)
\end{aligned}
$$

$$p_A = \Phi\left(\frac{|w^T x + b|}{\sigma\|w\|}\right)$$

$$
\begin{aligned}
R &= \sigma\Phi^{-1}(p_A) \\
&= \frac{|w^T x + b|}{\|w\|}
\end{aligned}
$$



Neyman-Pearson Lemma (revisit)

- The “Best” rejection region
- Alpha and beta levels
- Type I and type II errors

The Neyman Pearson Lemma

Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a probability distribution with parameter θ. Then, if C is a critical region of size α and k is a constant such that:

$$\frac{L(\theta_0)}{L(\theta_\alpha)} \leq k \text{ inside the critical region } C$$

and:

$$\frac{L(\theta_0)}{L(\theta_\alpha)} \geq k \text{ outside the critical region } C$$

then C is the best, that is, most powerful, critical region for testing the simple null hypothesis $H_0: \theta = \theta_0$ against the simple alternative hypothesis $H_A: \theta = \theta_a$.

Lemma 3 (Neyman-Pearson). *Let X and Y be random variables in $\mathbb{R}^d$ with densities μ_X and μ_Y. Let $h : \mathbb{R}^d \to \{0, 1\}$ be a random or deterministic function. Then:*

1. *If $S = \left\{ z \in \mathbb{R}^d : \frac{\mu_Y(z)}{\mu_X(z)} \leq t \right\}$ for some $t > 0$ and $\mathbb{P}(h(X) = 1) \geq \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \geq \mathbb{P}(Y \in S)$.*
2. *If $S = \left\{ z \in \mathbb{R}^d : \frac{\mu_Y(z)}{\mu_X(z)} \geq t \right\}$ for some $t > 0$ and $\mathbb{P}(h(X) = 1) \leq \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y) = 1) \leq \mathbb{P}(Y \in S)$.*

Takeaways

- “Smoothed” classifiers can improve the consistency of nearby regions for a given instance
- The best test from Neyman-pearson provides tight bound for the certified robustness
- There are many variations of certified robustness via randomized smoothing

Provable Defenses Against Adversarial Examples via the convex outer adversarial polytope

- A method to learn deep ReLU-based classifiers which are provably robust against norm bounded adversarial perturbations
- Consider a convex outer approximation of the set of activations reachable through a norm-bounded perturbation
- A robust optimization procedure that minimizes the worst case loss over the outer region (linear program)
- Execute a few more forward and backward passes through a modified network and achieve provable robustness to *any* norm-bounded adv

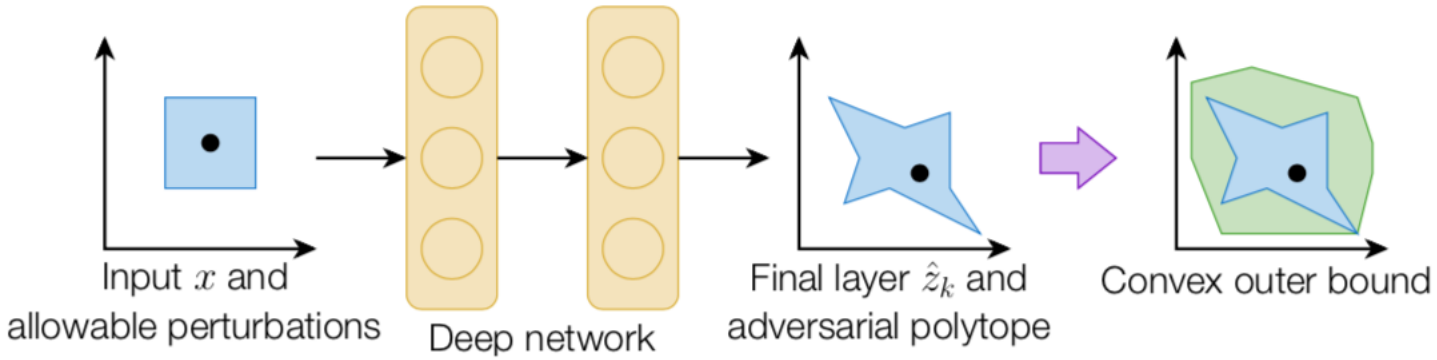
Adversarial polytope for deep ReLU networks

- Given a ReLU based network

$$\hat{z}_{i+1} = W_i z_i + b_i, \quad \text{for } i = 1, \ldots, k-1$$
$$z_i = \max\{\hat{z}_i, 0\}, \quad \text{for } i = 2, \ldots, k-1$$

- W_i represents a linear operator such as multiply or convolution

$$\mathcal{Z}_\epsilon(x) = \{f_\theta(x + \Delta) : \|\Delta\|_\infty \leq \epsilon\}$$

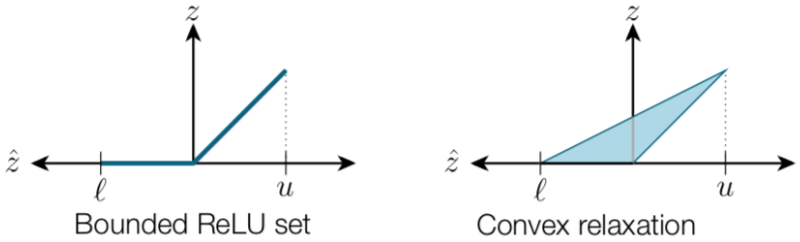


No point within this outer approximation exists that will change the class prediction of an example

Adversarial polytope for deep ReLU networks

- Linear relaxation of ReLU

$$z \geq 0,\ z \geq \hat{z},\ -u\hat{z} + (u-\ell)z \leq -u\ell.$$



- Robust guarantees via the convex outer adversarial polytope

$$\underset{\hat{z}_k}{\text{minimize}}\ \ (\hat{z}_k)_{y^\star} - (\hat{z}_k)_{y^{\text{targ}}} \equiv c^T \hat{z}_k$$

$$\text{subject to}\ \ \hat{z}_k \in \tilde{\mathcal{Z}}_\epsilon(x)$$

$\tilde{\mathcal{Z}}_\epsilon(x)$ Denotes the outer bound on the adversarial polytope from replacing the ReLU constraints

$c \equiv e_{y^\star} - e_{y^{\text{targ}}}$

- False positive? False negative?
  - +, 0

Challenges:
1. Solve the LP for each examples for each target is intractable;
2. How to compute l and u.

Efficient Optimization via the Dual Network

- Dual problem

minimize	$c^T x$	maximize	$-b^T \nu$
subject to	$Ax = b$	subject to	$A^T \nu + c \succeq 0$
	$x \succeq 0$		

Theorem 1. *The dual of (4) is of the form*

$$\begin{array}{ll} \underset{\alpha}{\text{maximize}} & J_\epsilon(x, g_\theta(c, \alpha)) \\ \text{subject to} & \alpha_{i,j} \in [0,1], \ \forall i, j \end{array} \tag{5}$$

where $J_\epsilon(x, \nu)$ is equal to

$$-\sum_{i=1}^{k-1} \nu_{i+1}^T b_i - x^T \hat{\nu}_1 - \epsilon \|\hat{\nu}_1\|_1 + \sum_{i=2}^{k-1} \sum_{j \in \mathcal{I}_i} \ell_{i,j} [\nu_{i,j}]_+ \tag{6}$$

and $g_\theta(c, \alpha)$ is a k layer feedforward neural network given by the equations

$$\begin{aligned} \nu_k &= -c \\ \hat{\nu}_i &= W_i^T \nu_{i+1}, \quad \text{for } i = k-1, \ldots, 1 \\ \nu_{i,j} &= \begin{cases} 0 & j \in \mathcal{I}_i^- \\ \hat{\nu}_{i,j} & j \in \mathcal{I}_i^+ \\ \frac{u_{i,j}}{u_{i,j} - \ell_{i,j}} [\hat{\nu}_{i,j}]_+ - \alpha_{i,j} [\hat{\nu}_{i,j}]_- & j \in \mathcal{I}_i, \end{cases} \\ & \qquad \text{for } i = k-1, \ldots, 2 \end{aligned} \tag{7}$$

Any feasible dual solution provides a guaranteed lower bound on the solution of the primal

Efficient Robust Optimization

- Standard robust optimization

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{N} \max_{\|\Delta\|_\infty \leq \epsilon} L(f_\theta(x_i + \Delta), y_i)$$

Theorem 2. *Let L be a monotonic loss function that satisfies Property 1. For any data point (x, y), and $\epsilon > 0$, the worst case adversarial loss from (11) can be upper bounded by*

$$\max_{\|\Delta\|_\infty \leq \epsilon} L(f_\theta(x + \Delta), y) \leq L(-J_\epsilon(x, g_\theta(e_y 1^T - I)), y),$$

- Distances to decision boundary

$$\underset{\epsilon}{\text{maximize}} \ \epsilon$$

$$\text{subject to} \ \ J_\epsilon(x, g_\theta(e_{f_\theta(x)} 1^T - I))_y \geq 0$$

Experiments

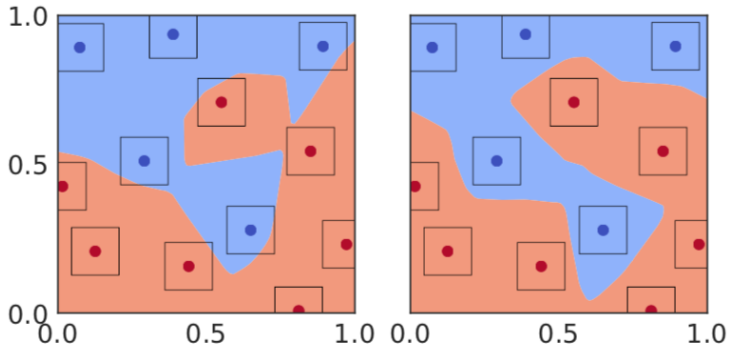


Figure 3. Illustration of classification boundaries resulting from standard training (left) and robust training (right) with ℓ_∞ balls of size $\epsilon = 0.08$ (shown in figure).

PROBLEM	ROBUST	ϵ	TEST ERROR	FGSM ERROR	PGD ERROR	ROBUST ERROR BOUND
MNIST	×	0.1	1.07%	50.01%	81.68%	100%
MNIST	√	0.1	1.80%	3.93%	4.11%	5.82%
FASHION-MNIST	×	0.1	9.36%	77.98%	81.85%	100%
FASHION-MNIST	√	0.1	21.73%	31.25%	31.63%	34.53%
HAR	×	0.05	4.95%	60.57%	63.82%	81.56%
HAR	√	0.05	7.80%	21.49%	21.52%	21.90%
SVHN	×	0.01	16.01%	62.21%	83.43%	100%
SVHN	√	0.01	20.38%	33.28%	33.74%	40.67%

Similar reading

- Certifying some distributional robustness with principled adversarial training

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)].$$

The Lagrangian relaxation for a fixed penalty

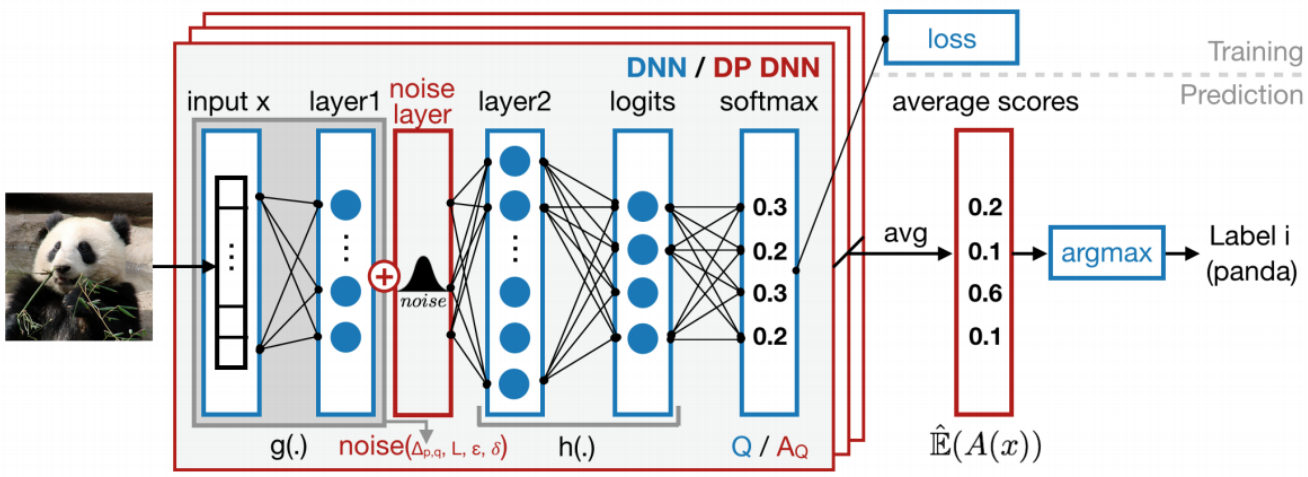
$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ F(\theta) := \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, P_0)\} = \mathbb{E}_{P_0}[\phi_\gamma(\theta; Z)] \right\}$$

$$\text{where } \phi_\gamma(\theta; z_0) := \sup_{z \in \mathcal{Z}} \{\ell(\theta; z) - \gamma c(z, z_0)\}.$$

For benign data, previous work obtain worse accuracy than this one

Differential privacy VS. robustness

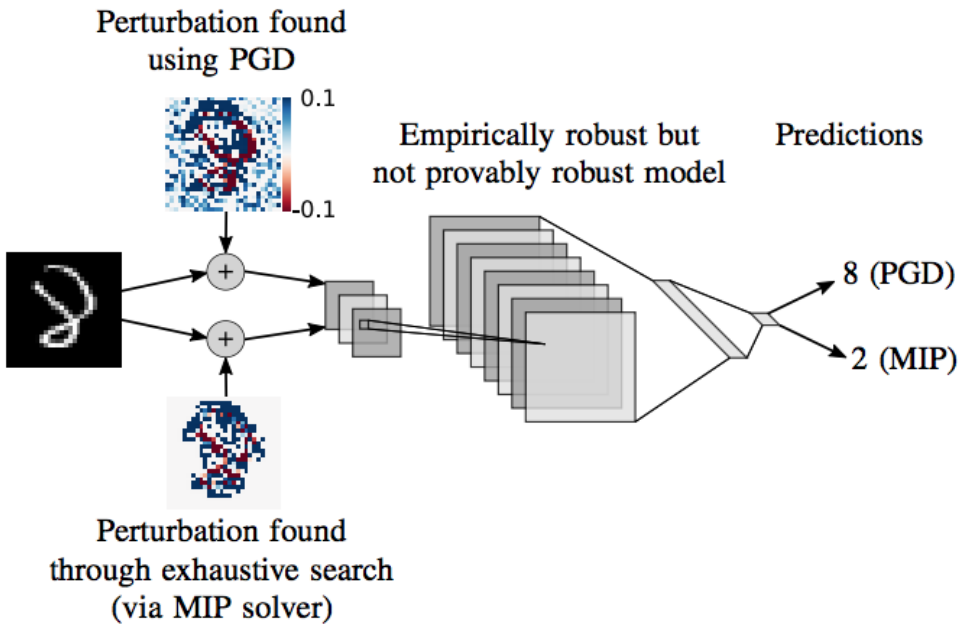
- Certified Robustness to Adversarial Examples with Differential Privacy



Takeaways

- Leveraging dual of the primal constrained optimization to provide provable robustness guarantee
- Linear relaxation would lead to loose robustness bound

On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models



Robustness to the projected gradient descent (PGD) attack is not a true measure of robustness (even for small convolutional neural networks). Given a seemingly robust neural network, the worst-case perturbation of size 0.1 found using 200 PGD iterations and 10 random restarts (shown at the top) is correctly classified as an “eight”. However, a worst case perturbation classified as a “two” can be found through exhaustive search (shown at the bottom)

On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models

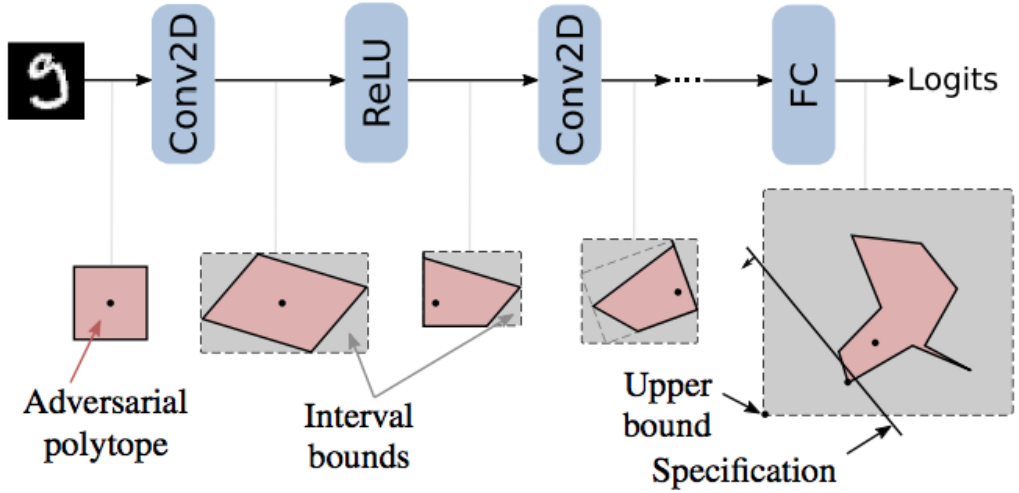


Illustration of interval bound propagation. From the left, the adversarial polytope (illustrated in 2D for clarity) of the nominal image of a “nine” (in red) is propagated through a convolutional network. At each layer, the polytope deforms itself until the last layer where it takes a complicated and non-convex shape in logit space. Interval bounds (in gray) can be propagated similarly: after each layer the bounds are reshaped to be axis-aligned bounding boxes that always encompass the adversarial polytope. In logit space, it becomes easy to compute an upper bound on the worst case violation of the specification to verify