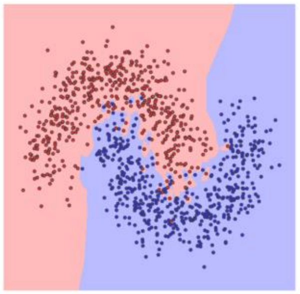
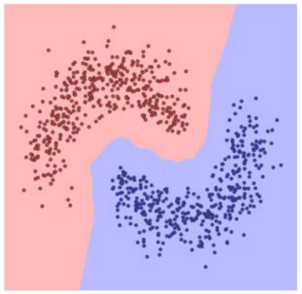



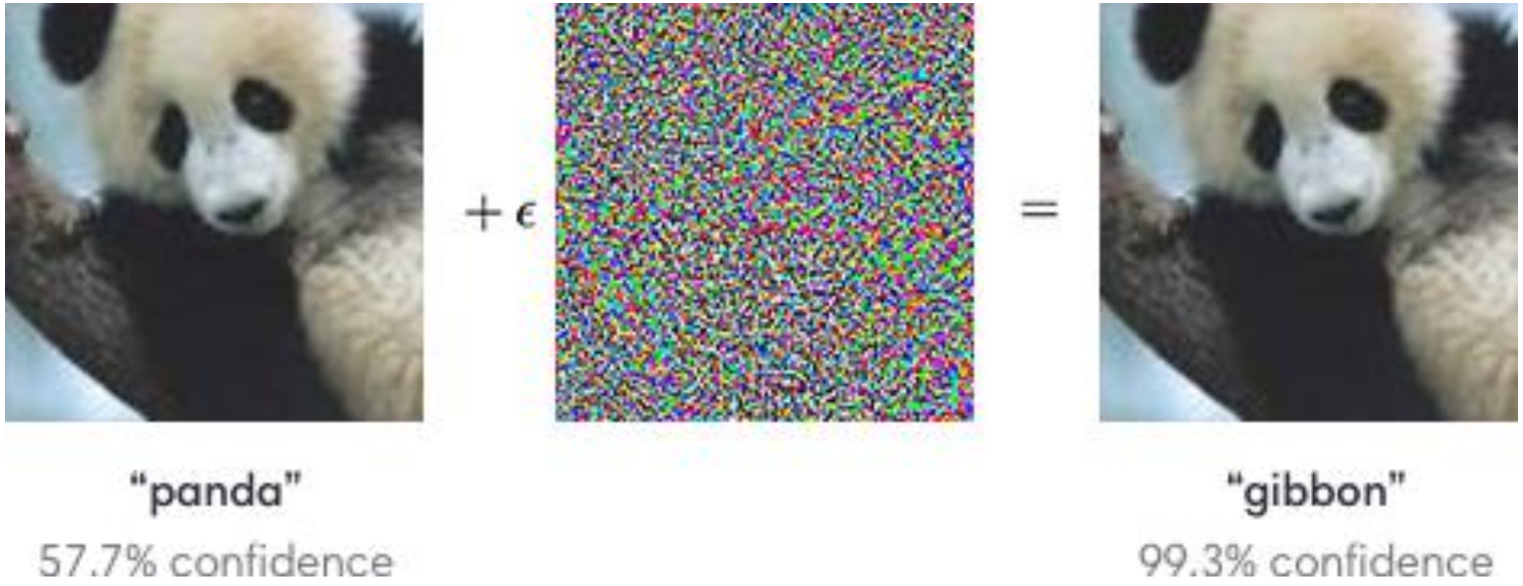
Adversarial Robustness From Well-Separated Data

Cyrus Rashtchian
UC San Diego

Joint with Yao-Yuan Yang, Yizhen Wang, Kamalika Chaudhuri **AISTATS 2020**

+ Hongyang Zhang, Ruslan Salakhutdinov **NeurIPS 2020**

Adversarial Examples



Lowd and Meek 2005

Szegedy et al 2013

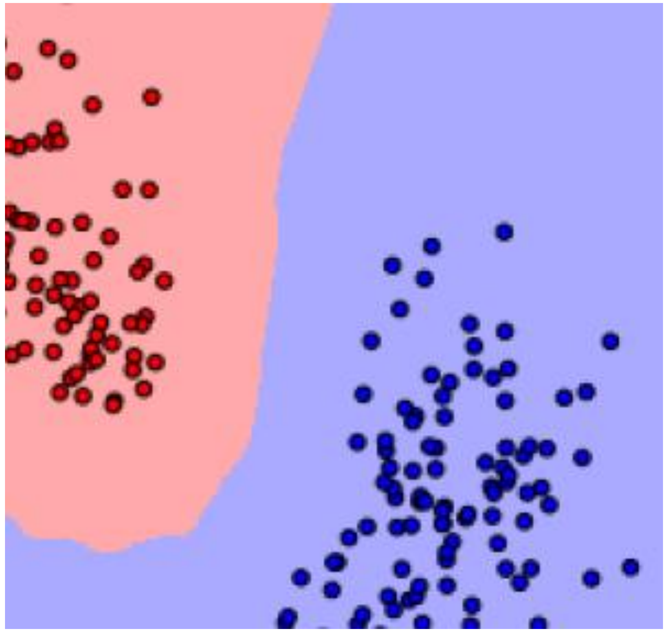
Papernot et al 2014,15,16

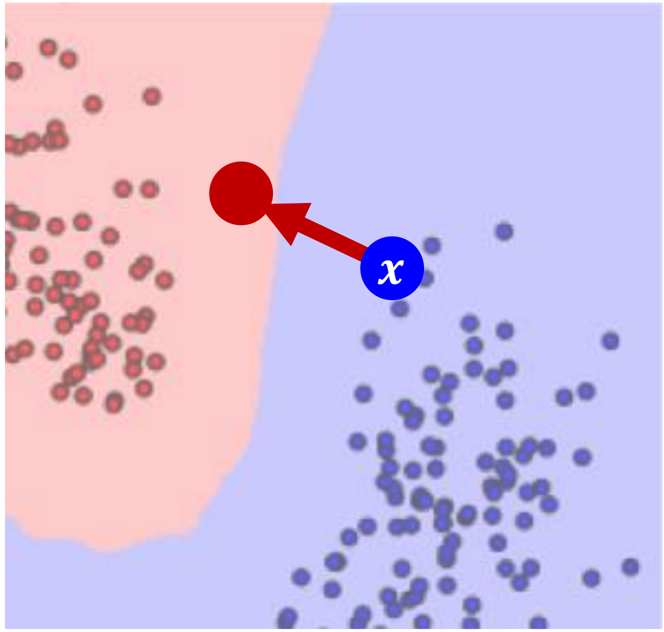
…

labradoodle or fried chicken

chihuahua or muffin

Karen Zack (aka @teenybiscuit)

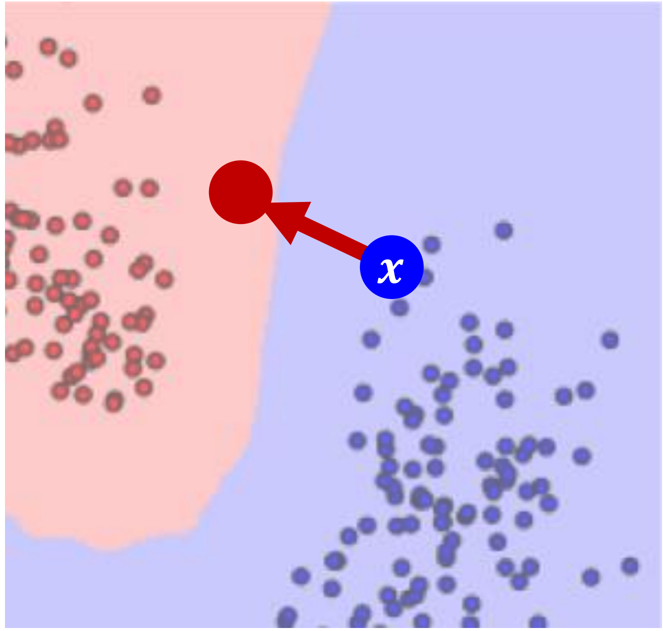



x

Definition (Adversarial example)

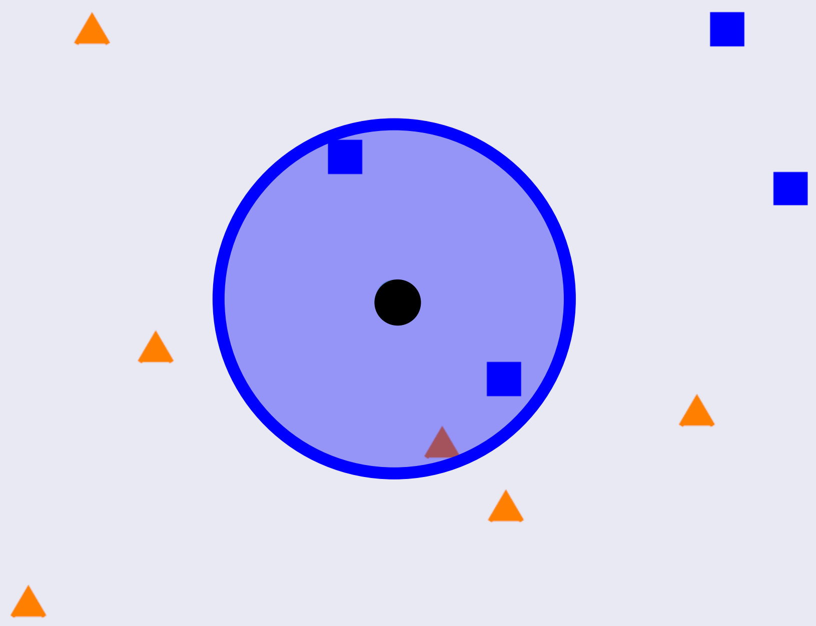
$\mathbf{x}_{adv}$ is an adversarial example of the target example $\mathbf{x}$ if and only if

$$\|\mathbf{x} - \mathbf{x}_{adv}\|_p \leq r \text{ and } f(\mathbf{x}) \neq f(\mathbf{x}_{adv})$$



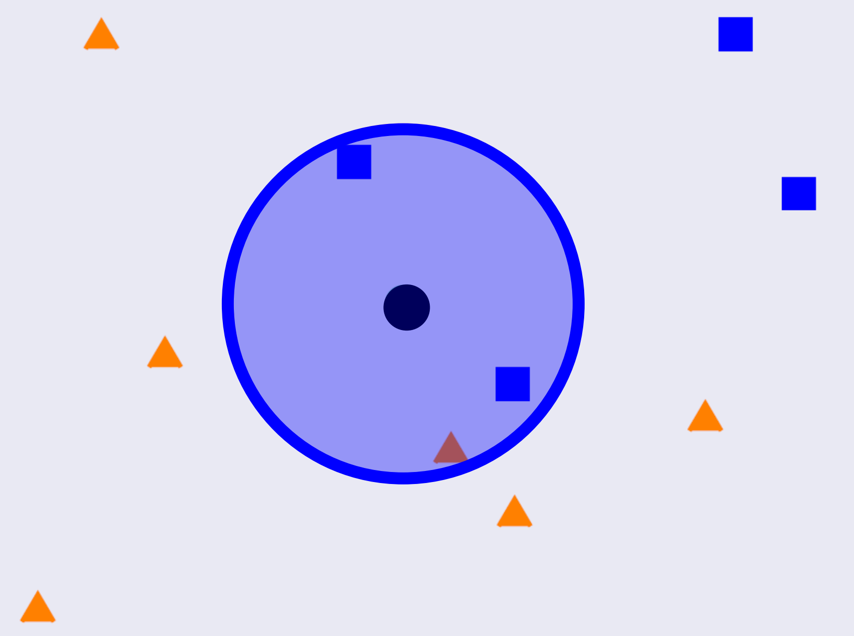
k nearest neighbor (k-NN)

take k closest training examples and output the majority label



k nearest neighbor (k-NN)

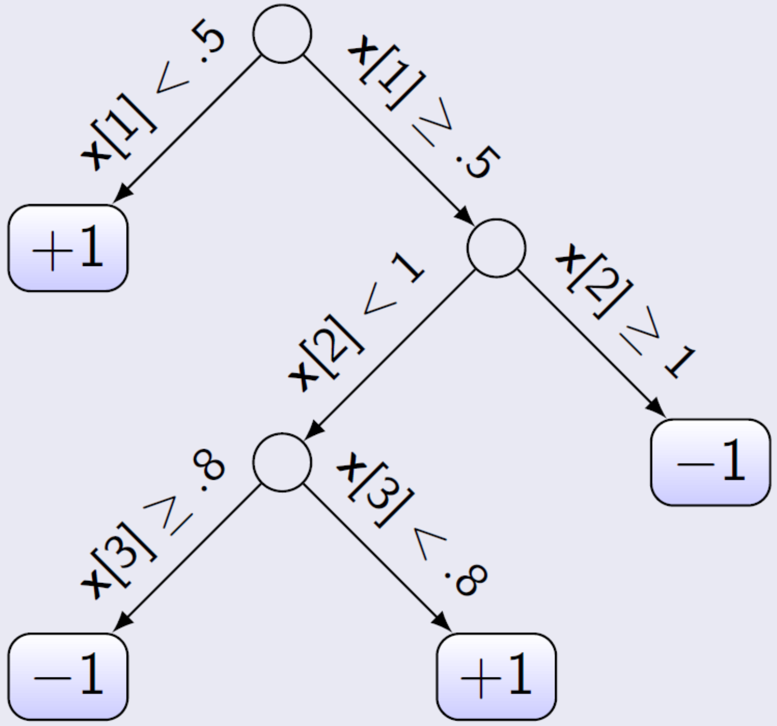
take k closest training examples and output the majority label



Decision tree and tree ensembles

recursively split the data

- common models: decision tree, random forest, gradient boosting trees, etc.



Optimal attack

$$\min_{\mathbf{x}_{adv}} \|\mathbf{x} - \mathbf{x}_{adv}\|_p \quad \text{s.t.} \quad f(\mathbf{x}) \neq f(\mathbf{x}_{adv})$$

Kantchelian, Tygar, Joseph. Evasion and Hardening of Tree Ensemble Classifiers. ICML 2016.
Wang, Jha, Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. ICML 2018.

Optimal attack

$$\min_{\mathbf{x}_{adv}} \|\mathbf{x} - \mathbf{x}_{adv}\|_p \quad \text{s.t.} \quad f(\mathbf{x}) \neq f(\mathbf{x}_{adv})$$

Kantchelian et al.

- proved the optimal attack on tree ensemble is NP-complete with increasing number of trees
- formulate the attack as a mixed integer linear program (MILP)

Kantchelian, Tygar, Joseph. Evasion and Hardening of Tree Ensemble Classifiers. ICML 2016.
Wang, Jha, Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. ICML 2018.

Optimal attack

$$\min_{\mathbf{x}_{adv}} \|\mathbf{x} - \mathbf{x}_{adv}\|_p \;\; \text{s.t.} \;\; f(\mathbf{x}) \neq f(\mathbf{x}_{adv})$$

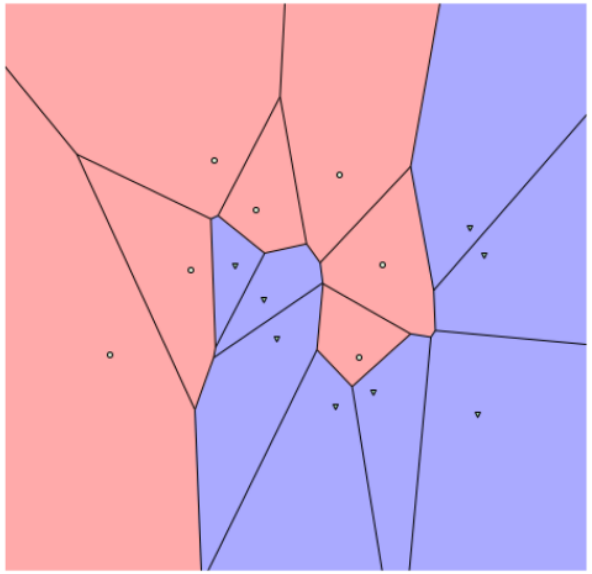
Kantchelian et al.

- proved the optimal attack on tree ensemble is NP-complete with increasing number of trees
- formulate the attack as a mixed integer linear program (MILP)

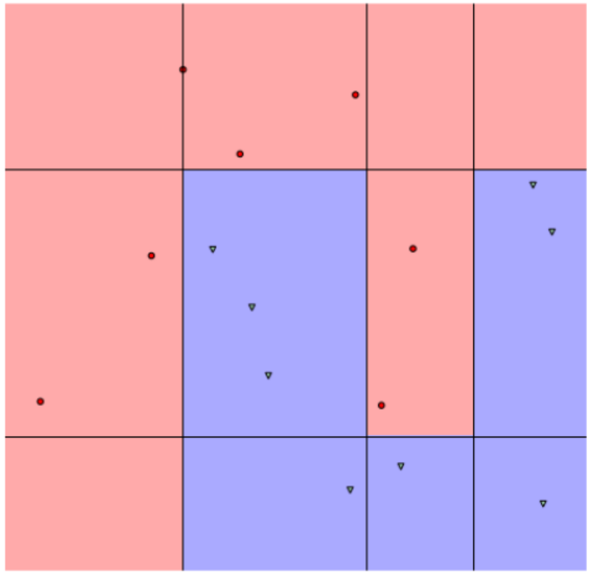
Prior attacks on non-parametrics

- model specific

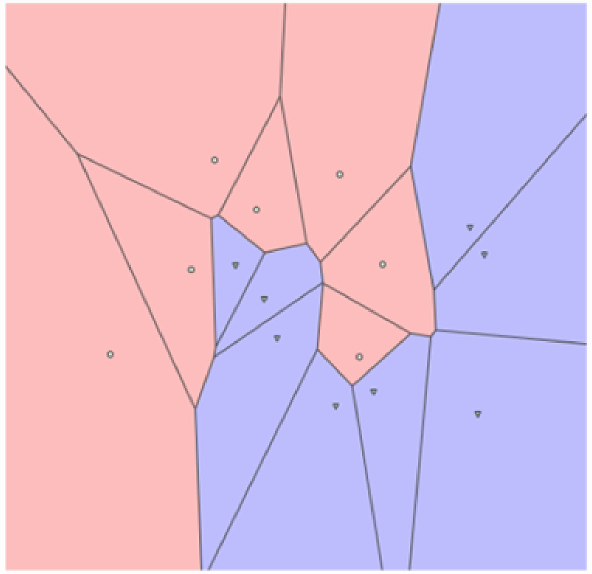
Kantchelian, Tygar, Joseph. Evasion and Hardening of Tree Ensemble Classifiers. ICML 2016.
Wang, Jha, Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. ICML 2018.



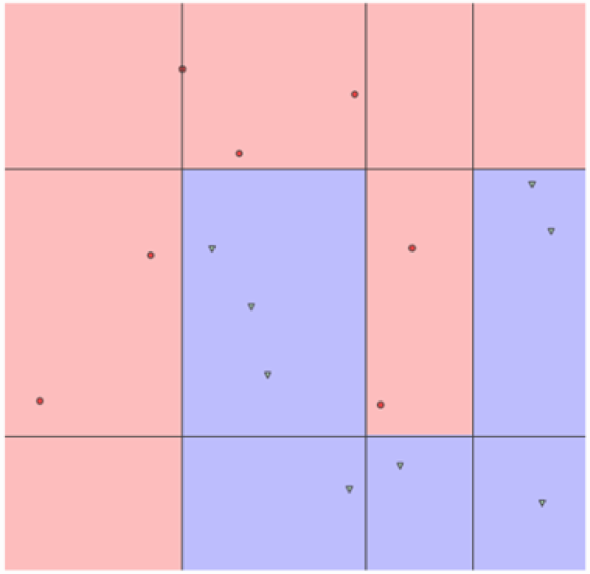
(a) 1-NN regions



(b) DT regions



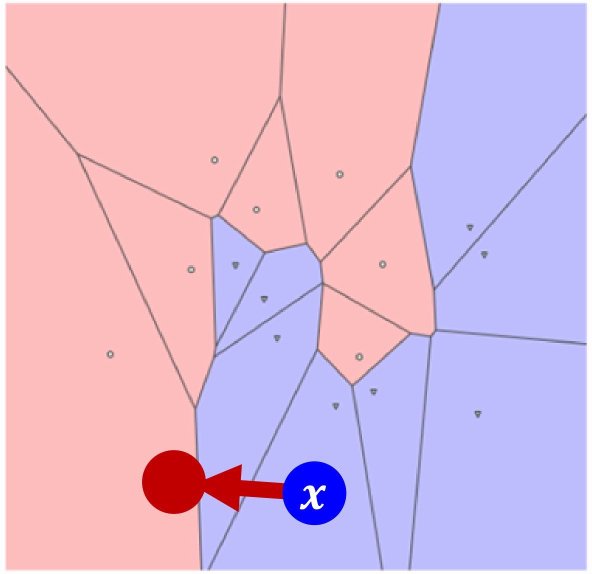
(a) 1-NN regions



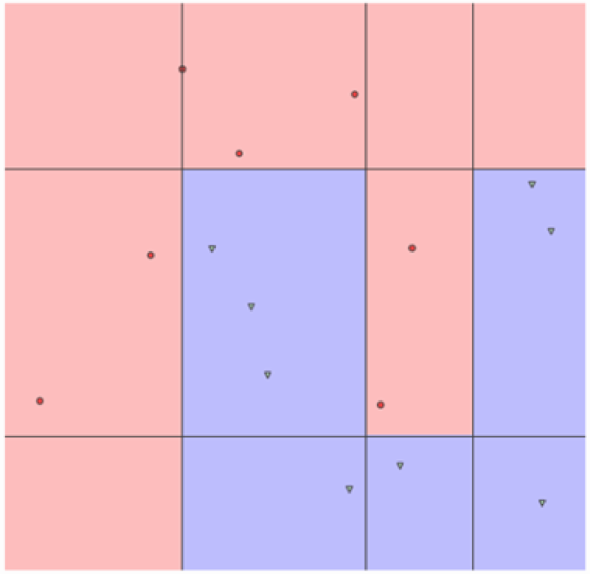
(b) DT regions

Definition ((s, m)-decomposition)

The partition of $\mathcal{R}^d$ into convex regions $P_1, \ldots, P_s$ s.t. each P_i can be described by at most m linear constraints.



(a) 1-NN regions



(b) DT regions

Definition ((*s*, *m*)-*decomposition*)

The partition of $\mathcal{R}^d$ into convex regions $P_1, \ldots, P_s$ s.t. each P_i can be described by at most m linear constraints.

Region-Based Attack

$$\min_{i:f(\mathbf{x})\neq y_i} \min_{\mathbf{x}_{adv}\in P_i} \|\mathbf{x}-\mathbf{x}_{adv}\|_p$$

- outer min: iterate through all regions
- inner min: LP for $p=1,\infty$ and QP for $p=2$

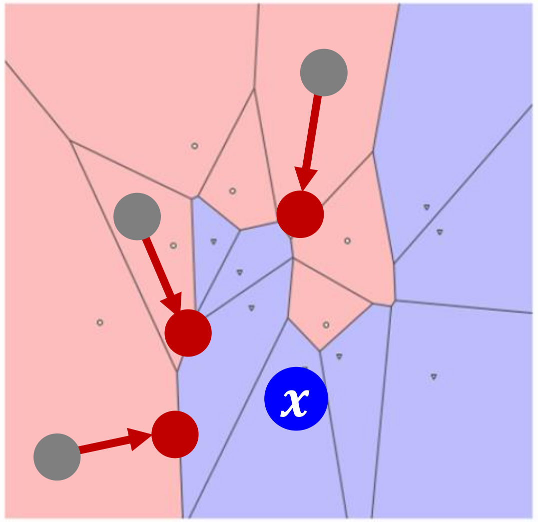
Definition ((*s*, *m*)-*decomposition*)

The partition of $\mathcal{R}^d$ into convex regions $P_1,\ldots,P_s$ s.t. each P_i can be described by at most m linear constraints.

Region-Based Attack

$$\min_{i:f(\mathbf{x})\neq y_i} \min_{\mathbf{x}_{adv}\in P_i} \|\mathbf{x}-\mathbf{x}_{adv}\|_p$$

- outer min: iterate through all regions
- inner min: LP for $p=1,\infty$ and QP for $p=2$

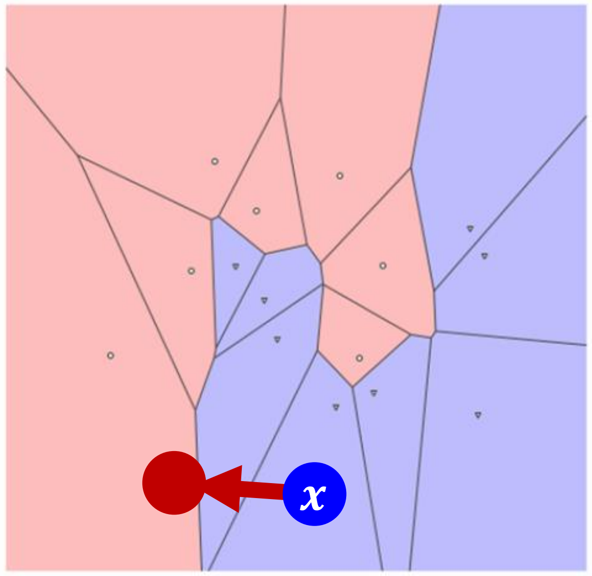


1-NN regions

RBA-Exact:
optimal adversarial example

RBA-Approx:
only consider subset of 50-100 regions containing training pts

Comparing Attacks



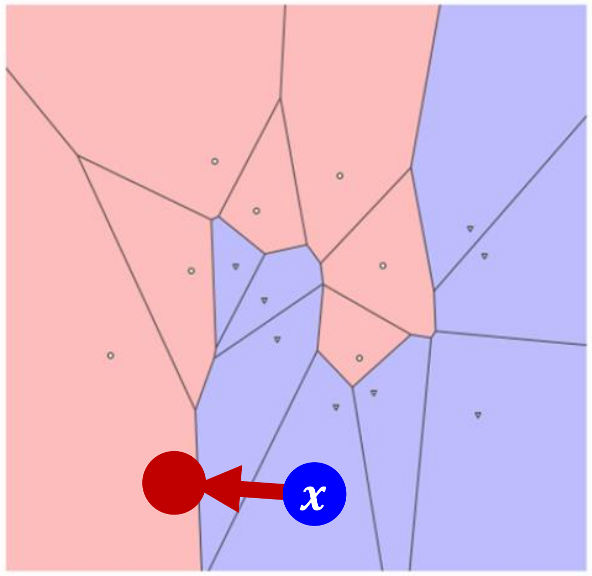
1-NN regions

Measure **distance to closest adversarial example** for examples in the test set

Lower == better attack

Comparing Attacks

	1-NN				3-NN			
	Direct	BBox	Kernel	RBA-Exact	Direct	BBox	Kernel	RBA-Approx



1-NN regions

Measure **distance to closest adversarial example** for examples in the test set*

*after PCA to 25 dim. for MNIST/F-MNIST

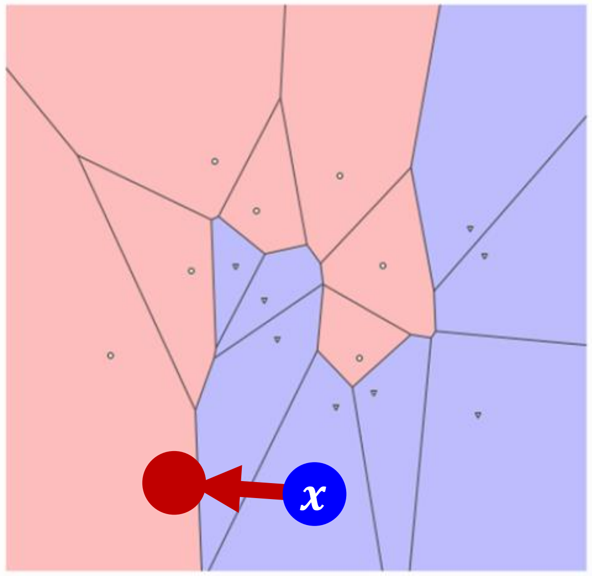
Lower == better attack

Cheng, Le, Chen, Yi, Zhang, Hsieh. Query-efficient Hard-label Blackbox Attack: An Optimization-based Approach. ICLR 2019.

Papernot, McDaniel, Goodfellow. Transferability in Machine Learning: From Phenomena to Black-box Attacks Using Adversarial Samples. 2016

Comparing Attacks

	1-NN				3-NN			
	Direct	BBox	Kernel	RBA-Exact	Direct	BBox	Kernel	RBA-Approx



1-NN regions

Measure **distance to closest adversarial example** for examples in the test set*

*after PCA to 25 dim. for MNIST/F-MNIST

Lower == better attack

Cheng, Le, Chen, Yi, Zhang, Hsieh. Query-efficient Hard-label Blackbox Attack: An Optimization-based Approach. ICLR 2019.
Papernot, McDaniel, Goodfellow. Transferability in Machine Learning: From Phenomena to Black-box Attacks Using Adversarial Samples. 2016

Comparing Attacks

	1-NN				3-NN			
	Direct	BBox	Kernel	RBA-Exact	Direct	BBox	Kernel	RBA-Approx
australian	.442	.336	.379	**.151**	.719	.391	.464	**.278**
cancer	.223	.364	.358	**.137**	.329	.376	.394	**.204**
covtype	.320	.207	.271	**.076**	.443	.265	.271	**.120**
diabetes	.074	.112	.165	**.035**	.130	.143	.191	**.078**
f-mnist06	.259	.162	.187	**.034**	.233	.184	.213	**.064**
f-mnist35	.354	.269	.288	**.089**	.355	.279	.295	**.111**
fourclass	.109	.124	.137	**.090**	.101	.113	.134	**.096**
halfmoon	.070	.129	.102	**.059**	.105	.132	.115	**.096**
mnist17	.330	.260	.239	**.079**	.302	.264	.247	**.098**

Cheng, Le, Chen, Yi, Zhang, Hsieh. Query-efficient Hard-label Blackbox Attack: An Optimization-based Approach. ICLR 2019.
Papernot, McDaniel, Goodfellow. Transferability in Machine Learning: From Phenomena to Black-box Attacks Using Adversarial Samples. 2016

Comparing Attacks

	1-NN				3-NN			
	Direct	BBox	Kernel	RBA-Exact	Direct	BBox	Kernel	RBA-Approx
australian	.442	.336	.379	**.151**	.719	.391	.464	**.278**
cancer	.223	.364	.358	**.137**	.329	.376	.394	**.204**
covtype	.320	.207	.271	**.076**	.443	.265	.271	**.120**
diabetes	.074	.112	.165	**.035**	.130	.143	.191	**.078**
f-mnist06	.259	.162	.187	**.034**	.233	.184	.213	**.064**
f-mnist35	.354	.269	.288	**.089**	.355	.279	.295	**.111**
fourclass	.109	.124	.137	**.090**	.101	.113	.134	**.096**
halfmoon	.070	.129	.102	**.059**	.105	.132	.115	**.096**
mnist17	.330	.260	.239	**.079**	.302	.264	.247	**.098**

Our attacks are 2-3x better

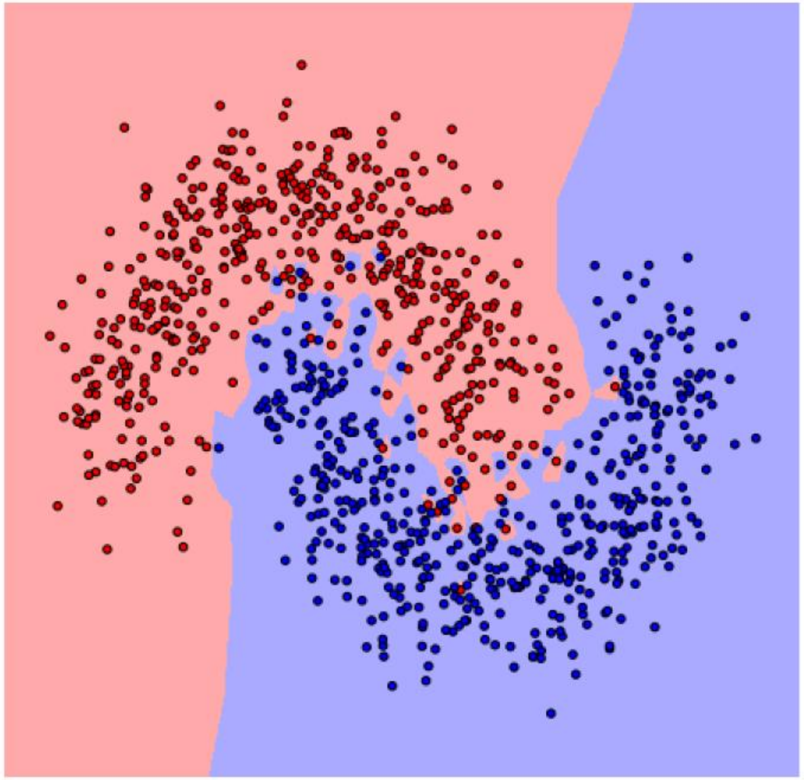
Cheng, Le, Chen, Yi, Zhang, Hsieh. Query-efficient Hard-label Blackbox Attack: An Optimization-based Approach. ICLR 2019.
Papernot, McDaniel, Goodfellow. Transferability in Machine Learning: From Phenomena to Black-box Attacks Using Adversarial Samples. 2016

Comparing Attacks cont.

	DT			RF	
	Papernot's	BBox	RBA-Exact	BBox	RBA-Approx
australian	.140	.139	**.070**	**.364**	.446
cancer	.459	.334	**.255**	.451	**.383**
covtype	.289	.117	**.070**	.256	**.219**
diabetes	.237	.133	**.085**	**.181**	.184
f-mnist06	.200	.182	**.114**	.222	**.199**
f-mnist35	.287	.168	**.112**	**.201**	.246
fourclass	.288	.197	**.137**	.159	**.133**
halfmoon	.098	.148	**.085**	.182	**.149**
mnist17	.236	.175	**.117**	**.237**	.244

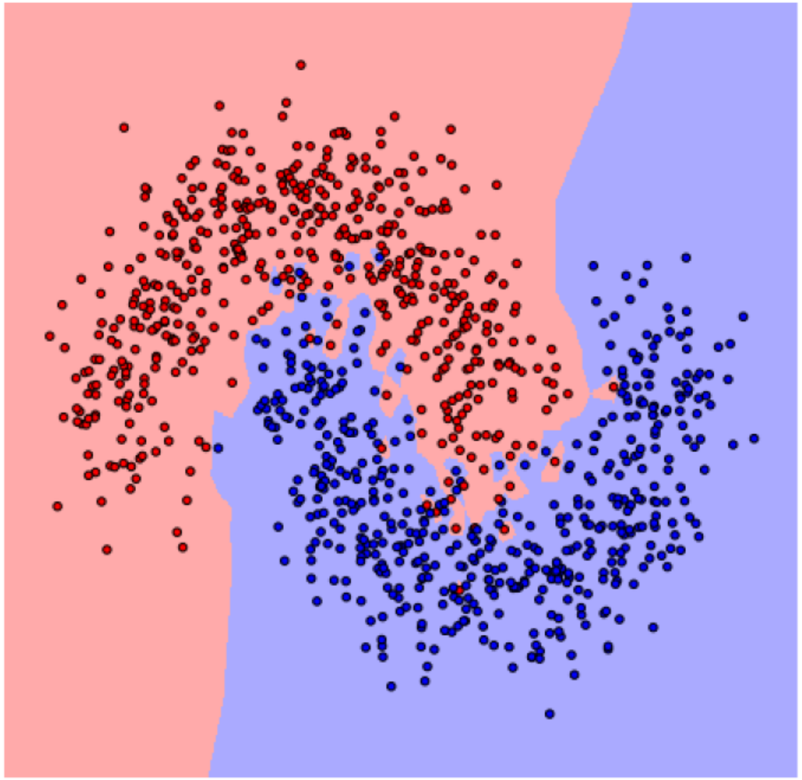
RFs are much harder to attack

Improving robustness via separation

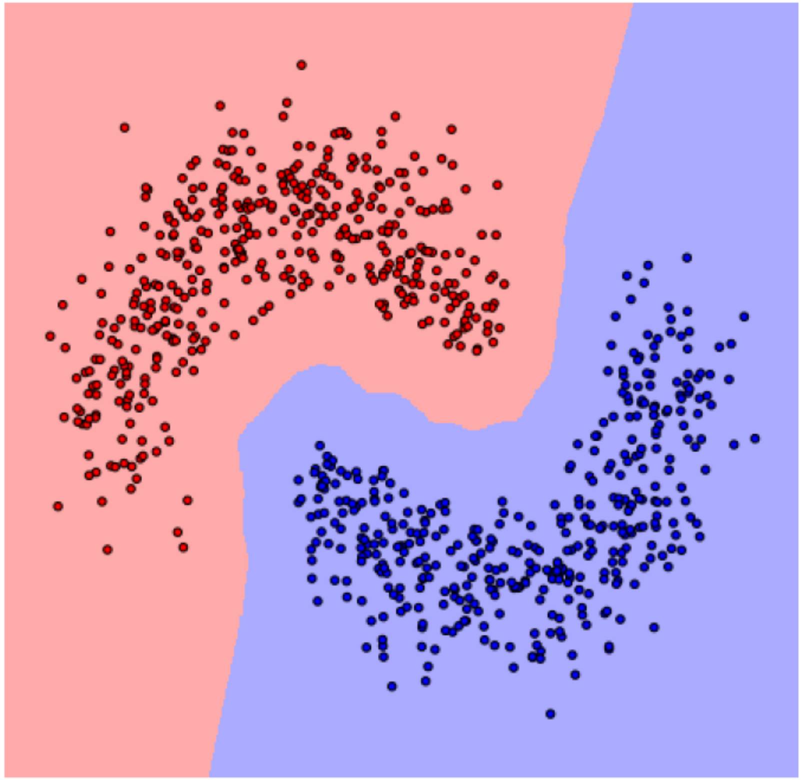


1-NN

Improving robustness via separation



1-NN

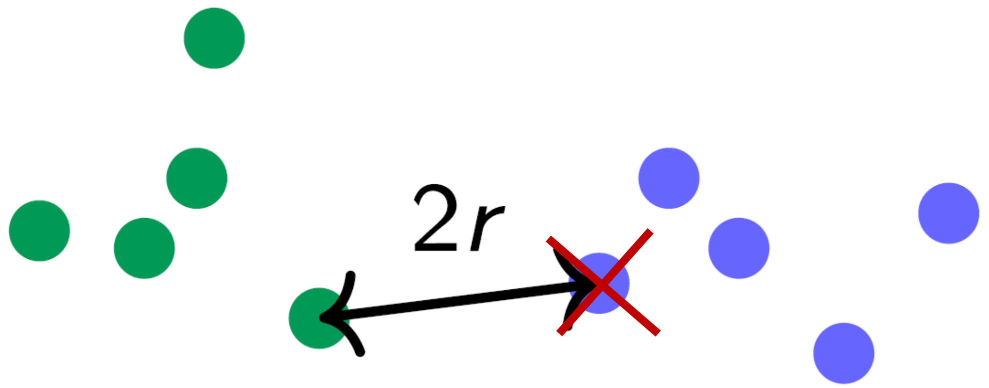


1-NN with separation (less overlap)

Adversarial Pruning

Defense strategy

1. remove minimum # of examples s.t. distance between opposite labeled examples $\geq 2r$
2. learn non-parametric classifier

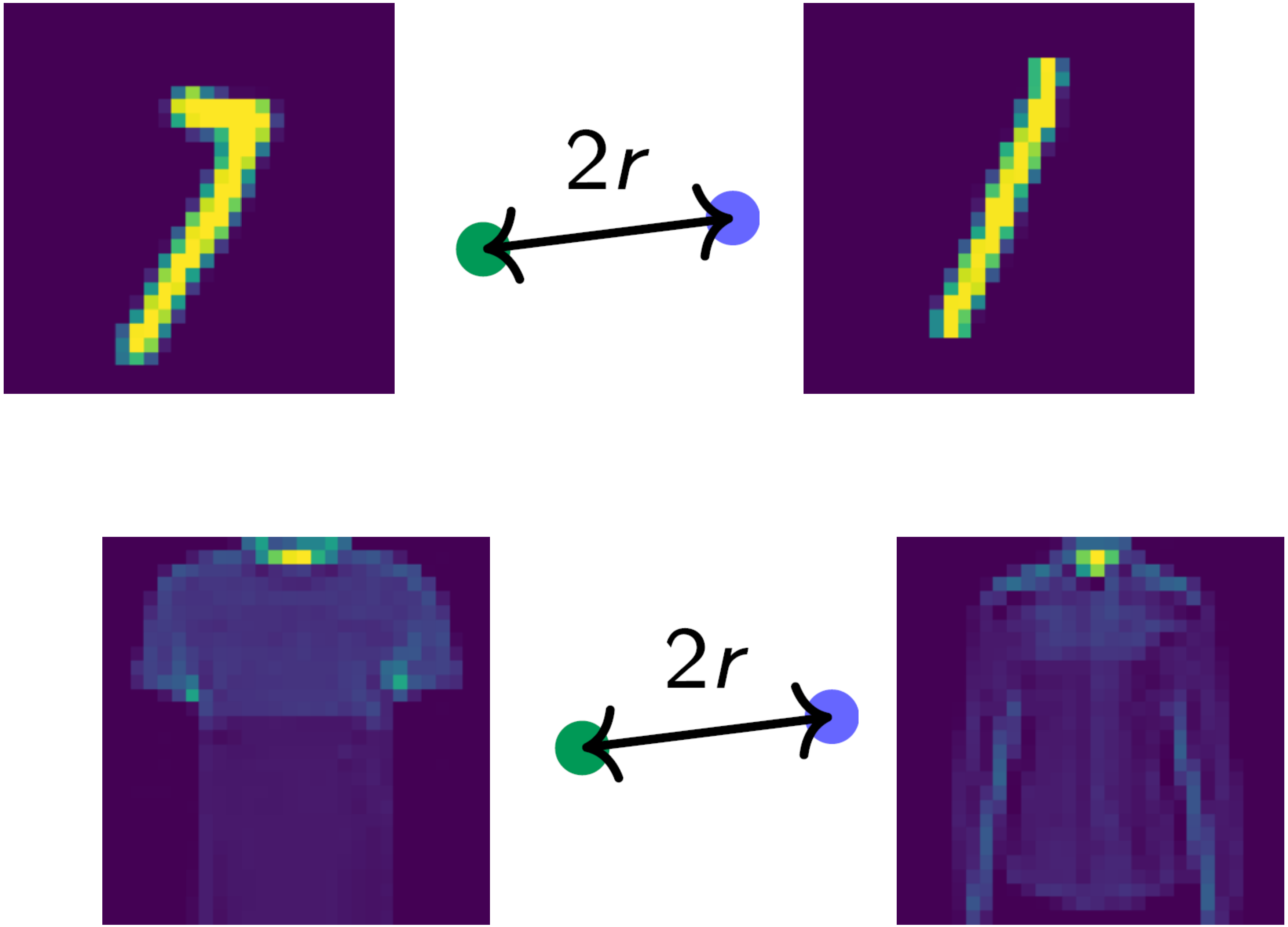


Computing Pruned Dataset

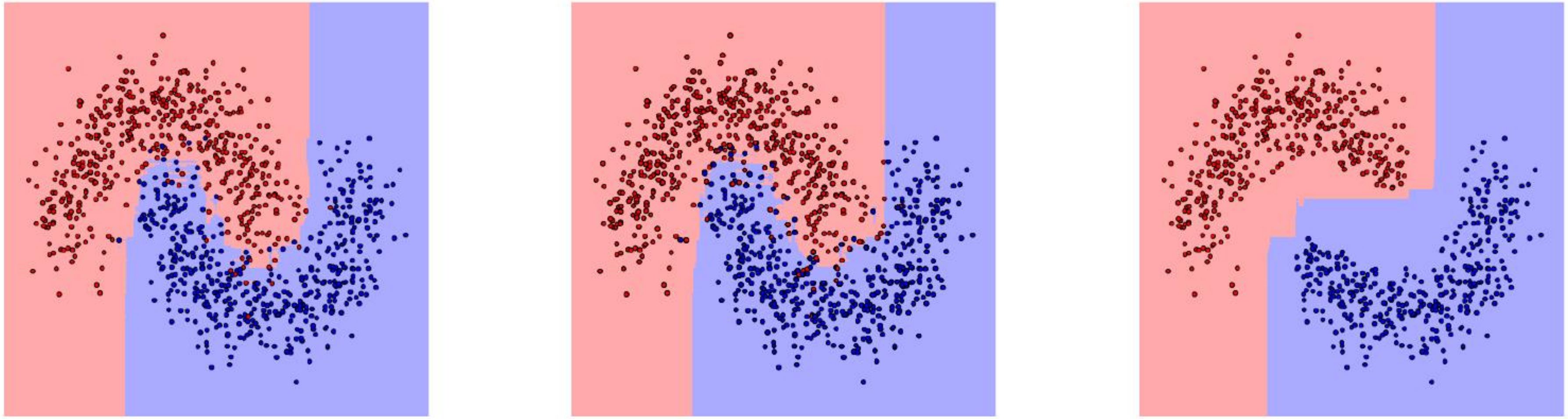
Bipartite maximum matching via Hopcroft-Karp algorithm (1973)

Graph n vertices and m edges, running time $O(m\sqrt{n})$

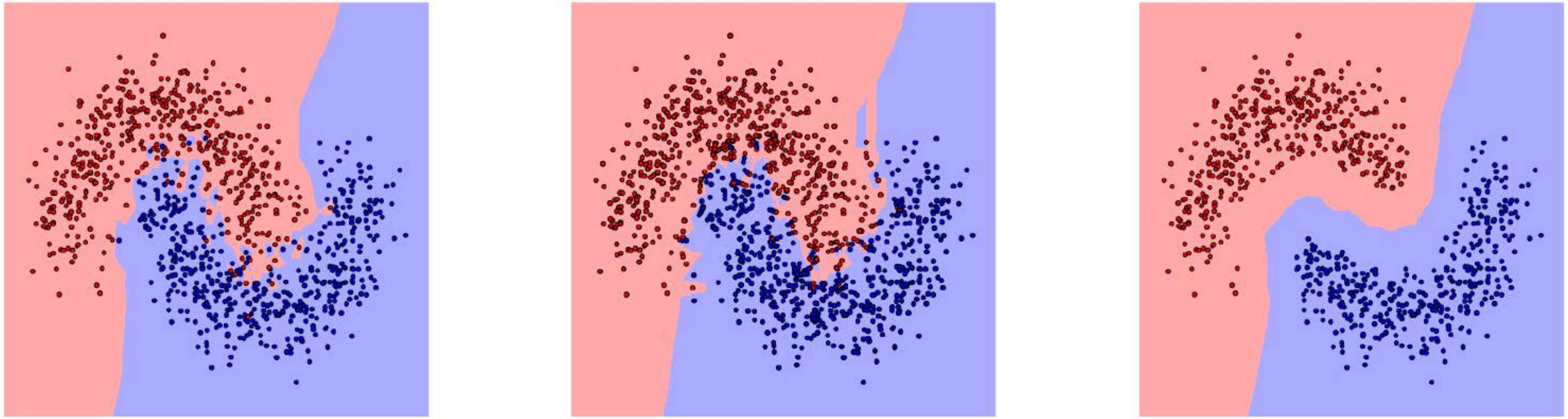
Gottlieb, Kontorovich, Krauthgamer. Efficient Classification for Metric Data. IEEE Transactions on Information Theory, 2014.
Kontorovich, Weiss. A Bayes Consistent 1-NN Classifier. AISTATS 2015



Distance in feature space after PCA to 25 dimensions



RF　　RF with AT　　RF with AP



1-NN　　1-NN with AT　　1-NN with AP

Evaluating defenses

	1-NN			3-NN		DT			RF		
	AT	Wang's	AP	AT	AP	AT	RS	AP	AT	RS	AP

Chen, Zhang, Boning, Hsieh. Robust Decision Trees Against Adversarial Examples. ICML 2018.
Wang, Jha, Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. ICML 2018.

Evaluating defenses

	1-NN			3-NN		DT			RF		
	AT	Wang's	AP	AT	AP	AT	RS	AP	AT	RS	AP

$$\text{defscore} = \frac{\textbf{defended dist. to adv. example}}{\textbf{undefended dist. to adv. example}}$$

average over test set
restrict to correctly classified (normalize accuracy)

Evaluating defenses

Baseline: Chen et al ICML'19
RS = Robust splitting

	1-NN			3-NN		DT			RF		
	AT	Wang's	AP	AT	AP	AT	RS	AP	AT	RS	AP

$$\text{defscore} = \frac{\textbf{defended dist. to adv. example}}{\textbf{undefended dist. to adv. example}}$$

average over test set
restrict to correctly classified (normalize accuracy)

Evaluating defenses

Baseline: Chen et al ICML'19
RS = Robust splitting

	1-NN			3-NN		DT			RF	
AT	Wang's	AP	AT	AP	AT	RS	AP	AT	RS	AP

$$\text{defscore} = \frac{\text{defended dist. to adv. example}}{\text{undefended dist. to adv. example}}$$

average over test set
restrict to correctly classified (normalize accuracy)

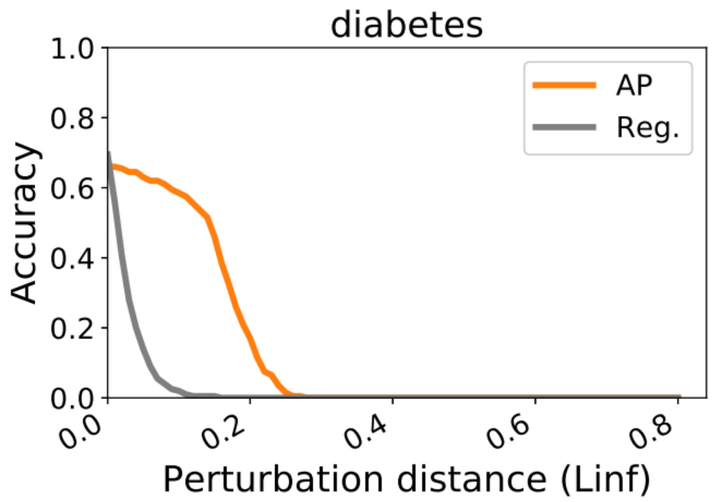
Evaluating defenses

Baseline: Chen et al ICML'19
RS = Robust splitting

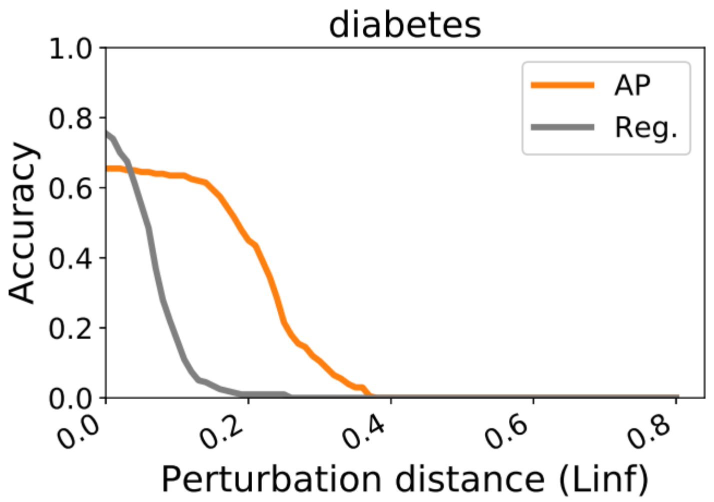
	1-NN			3-NN		DT			RF		
	AT	Wang's	AP	AT	AP	AT	RS	AP	AT	RS	AP
australian	0.64	**1.65**	**1.65**	0.68	**1.20**	2.36	**5.86**	2.37	1.07	**1.12**	1.04
cancer	0.82	1.05	**1.41**	1.06	**1.39**	0.85	1.09	**1.19**	0.87	**1.54**	1.26
covtype	0.61	**3.17**	**3.17**	0.81	**2.55**	1.07	2.90	**4.84**	0.93	1.59	**2.10**
diabetes	0.83	**4.69**	**4.69**	0.87	**2.97**	0.93	1.53	**2.22**	1.19	1.25	**2.22**
f-mnist06	0.94	2.09	**2.12**	0.86	**1.47**	0.82	**3.91**	1.85	0.97	1.17	**1.81**
f-mnist35	0.80	1.02	**1.08**	0.77	**1.05**	1.11	**2.64**	2.07	0.90	1.23	**1.32**
fourclass	0.93	**3.09**	**3.09**	0.89	**3.09**	1.06	1.23	**3.04**	1.03	1.92	**3.59**
halfmoon	1.03	1.98	**2.73**	0.93	**1.92**	1.54	1.98	**2.58**	1.04	1.01	**1.82**
mnist17	0.78	1.01	**1.20**	0.81	**1.13**	1.14	**2.91**	1.54	0.93	1.11	**1.29**

Higher == better defense

Chen, Zhang, Boning, Hsieh. Robust Decision Trees Against Adversarial Examples. ICML 2018.
Wang, Jha, Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. ICML 2018.

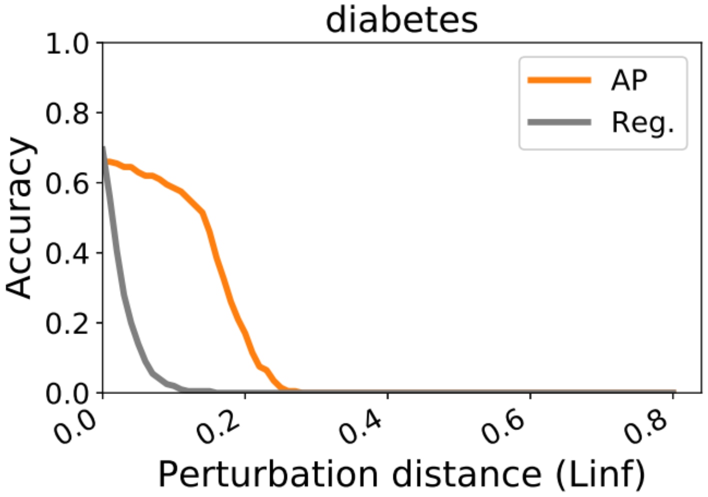


(e) 1-NN

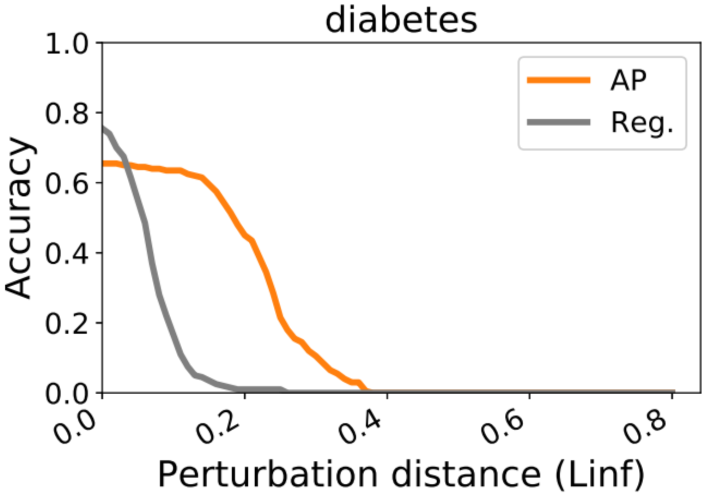


(f) 3-NN

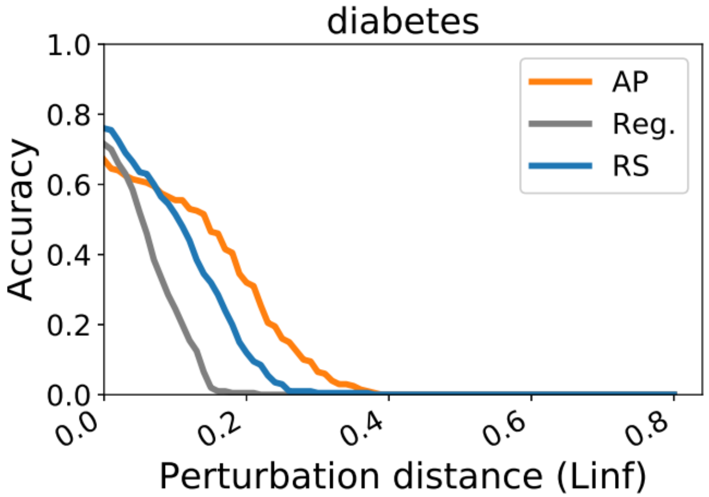
Our defense (AP) increases necessary perturbation distance



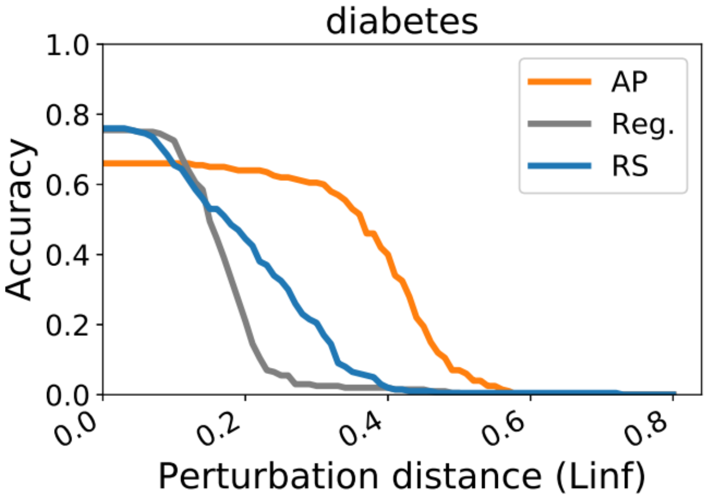
(e) 1-NN



(f) 3-NN



(g) Decision tree



(h) Random forest

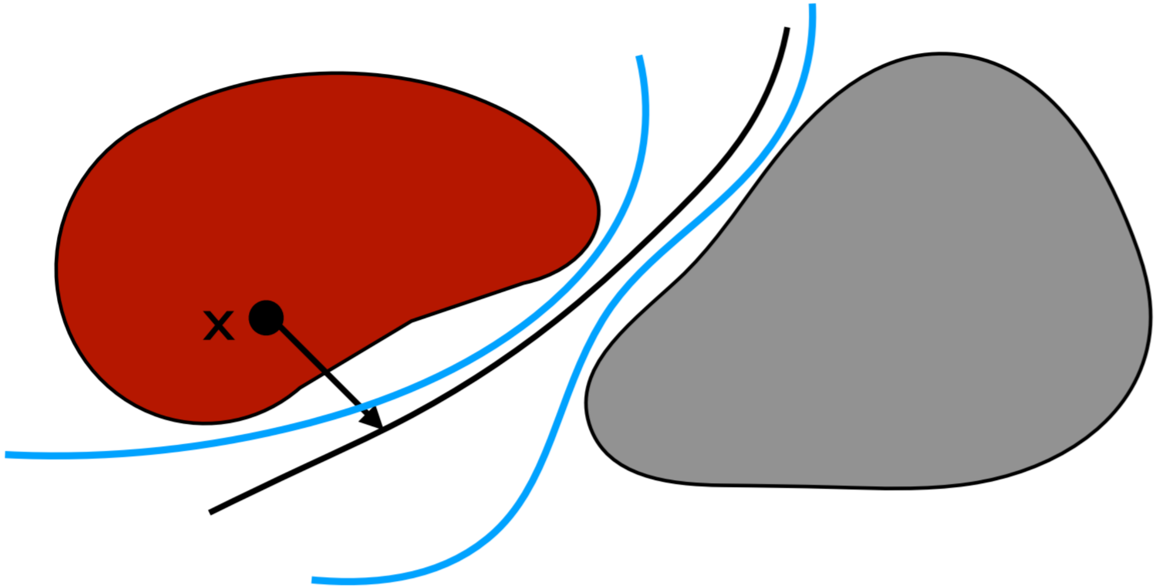
Our defense (AP) increases necessary perturbation distance

Downside:
Reduces accuracy ☹

Theoretical justification:

Robust analogue of Bayes Optimal

Robust analogue of Bayes Optimal

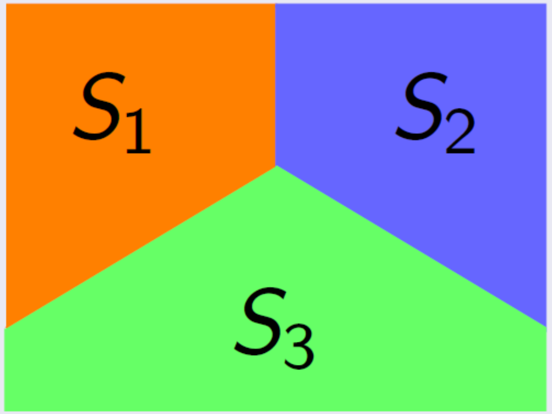


Bayes optimal undefined outside distribution

r-optimal = classifier that maximizes accuracy at points that have robustness radius at least r

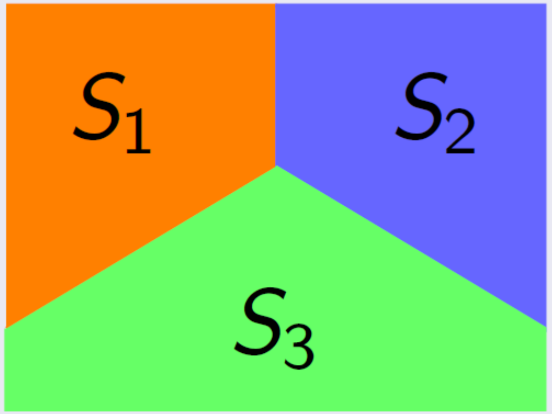
Bayes-optimal classifier

$$\max_{S_1,\ldots,S_c} \sum_{j=1}^{c} \int_{\mathbf{x}\in S_j} pr(y = j \mid \mathbf{x}) d\mu$$



Bayes-optimal classifier

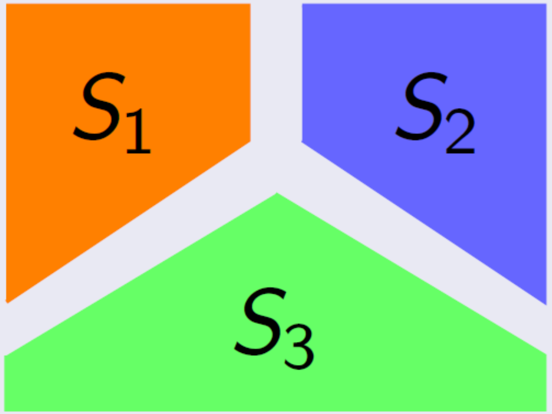
$$\max_{S_1,\ldots,S_c} \sum_{j=1}^{c} \int_{\mathbf{x}\in S_j} pr(y = j \mid \mathbf{x}) d\mu$$



r-Optimal classifier

$$\max_{S_1,\ldots,S_c} \sum_{j=1}^{c} \int_{\mathbf{x}\in S_j} pr(y = j \mid \mathbf{x}) d\mu$$

$$\text{s.t.} \quad d(S_j, S_{j'}) \geq 2r \quad \forall j \neq j'$$



Attack algorithm

- target model f
- target example $\mathbf{x}$
- attack budget r (defines "small")

attack algorithm $A(f, \mathbf{x}, r) \to \mathcal{R}^d$ returns an example under some attack budget constraint r

Definition (Astuteness $\text{ast}_\mu(A, f, r)$)

The accuracy after attack. Let μ be a distribution on $\mathcal{R}^d \times c$ and

$$\text{ast}_\mu(A, f, r) := \Pr_{(\mathbf{x}, y) \sim \mu} [f(\mathbf{x}) = f(A\ (f, \mathbf{x}, r)) \text{ and } f(\mathbf{x}) = y]$$

Attack algorithm

- target model f
- target example $\mathbf{x}$
- attack budget r (defines "small")

attack algorithm $A(f, \mathbf{x}, r) \to \mathcal{R}^d$ returns an example under some attack budget constraint r

Definition (Astuteness $\text{ast}_\mu(A, f, r)$)

The accuracy after attack. Let μ be a distribution on $\mathcal{R}^d \times c$ and

$$\text{ast}_\mu(A, f, r) := \Pr_{(\mathbf{x},y)\sim\mu}[f(\mathbf{x}) = f(A\ (f, \mathbf{x}, r)) \text{ and } f(\mathbf{x}) = y]$$

Theorem

r-Optimal classifier maximizes astuteness with attack radius r under μ.

$$f_{\text{ropt}} = \text{argmax}_f\ \text{ast}_\mu(f, r)$$

Follow-up Theory [Bhattacharjee, Chaudhuri 2020]

They prove that **Adversarial Pruning** + k-NN or kernel classifiers converges toward **optimally robust** and accurate classifiers (under certain conditions)

Definition (Astuteness $\mathrm{ast}_\mu(A, f, r)$)

The accuracy after attack. Let μ be a distribution on $\mathcal{R}^d \times c$ and

$$\mathrm{ast}_\mu(A, f, r) := \Pr_{(\mathbf{x},y)\sim\mu}[f(\mathbf{x}) = f(A\ (f, \mathbf{x}, r)) \text{ and } f(\mathbf{x}) = y]$$

Theorem

r-Optimal classifier maximizes astuteness with attack radius r under μ.

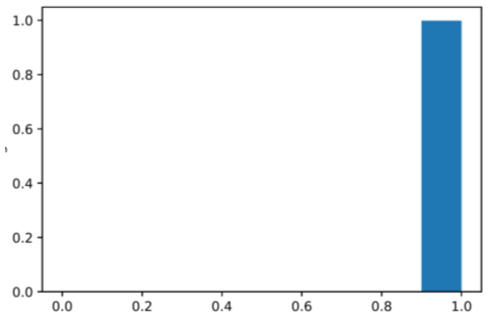
$$f_{\mathrm{ropt}} = \mathrm{argmax}_f\ \mathrm{ast}_\mu(f, r)$$

What about neural networks?

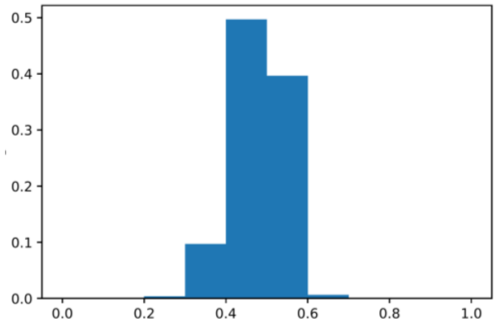
Can we get robustness + accuracy?

Separation of real datasets

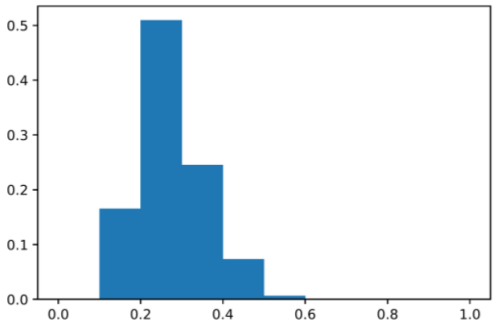
MNIST



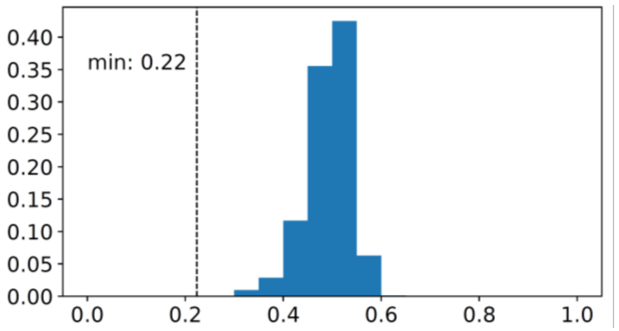
CIFAR-10



SVHN



ResImageNet

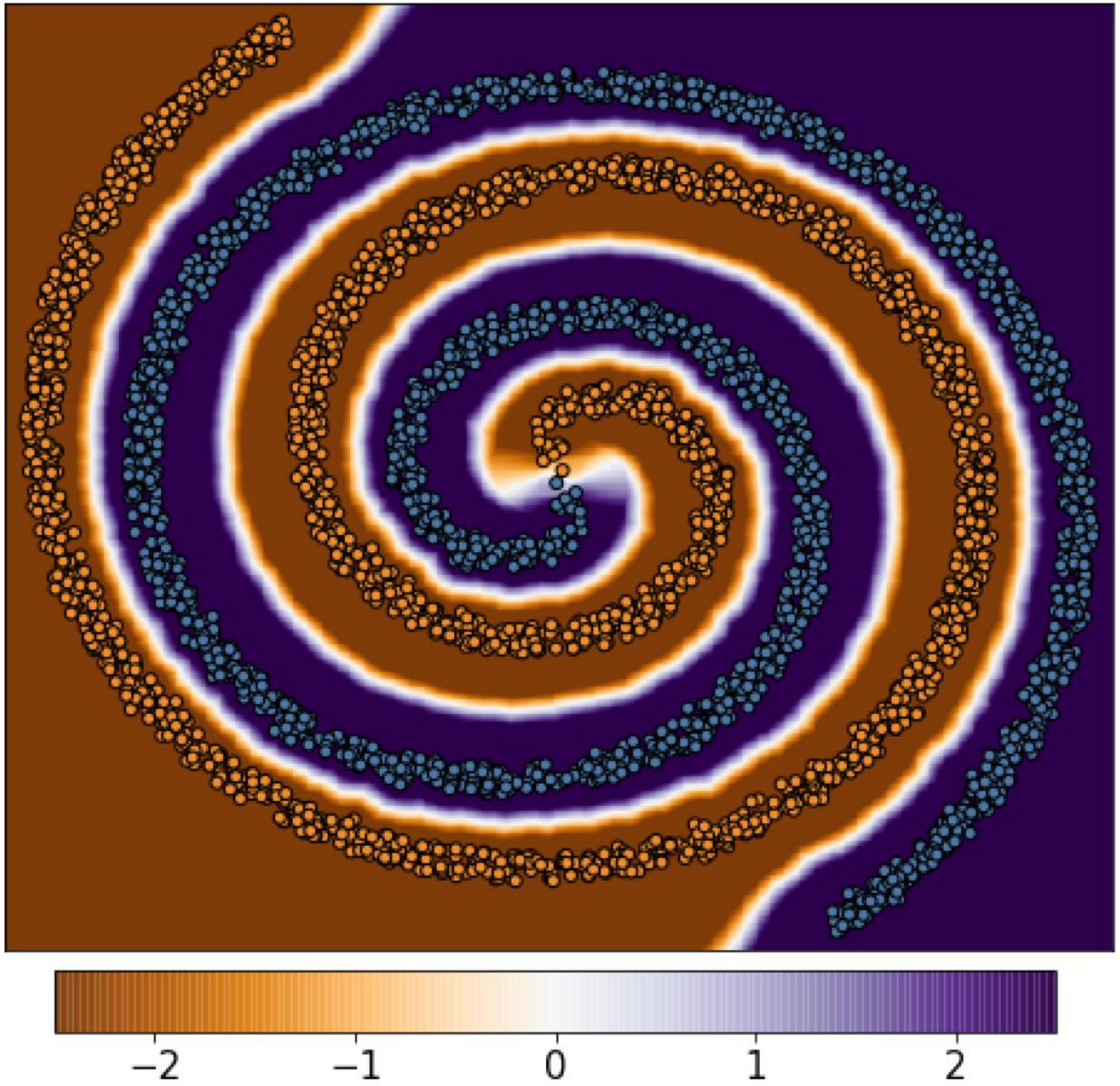


pairwise L_∞ distance

	typical perturbation distance	minimum Train-Train separation	minimum Test-Train separation
MNIST	0.1	0.737	0.812
CIFAR-10	0.031 (8/255)	0.212	0.220
SVHN	0.031 (8/255)	0.094	0.110
Restricted ImageNet	0.005	0.180	0.224

Thm.
For separated data there always exists a classifier that is

- Accurate
- Robust
- Locally Lipschitz



Thm.
For separated data there always exists a classifier that is

- Accurate
- Robust
- Locally Lipschitz

Perfect accuracy & robustness, at least in theory

Locally Lipschitz:

A function f is L-Locally Lipschitz in a radius r around x
if for all x' s.t. $d(x, x') \leq r$, we have $|f(x) - f(x')| < L \cdot d(x, x')$

Key idea behind all (provable) results for adversarial robustness

---

Hein, Andriushchenko. Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation. NeurIPS 2017.
Cohen, Rosenfeld, Kolter. Certified Adversarial Robustness via Randomized Smoothing. ICML 2019.
Salman, Yang, Li, Zhang, Zhang, Razenshteyn, Bubeck. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. NeurIPS 2019.

Perfect accuracy & robustness, at least in theory

Locally Lipschitz:

A function f is L-Locally Lipschitz in a radius r around x
if for all x' s.t. $d(x, x') \leq r$, we have $|f(x) - f(x')| < L \cdot d(x, x')$

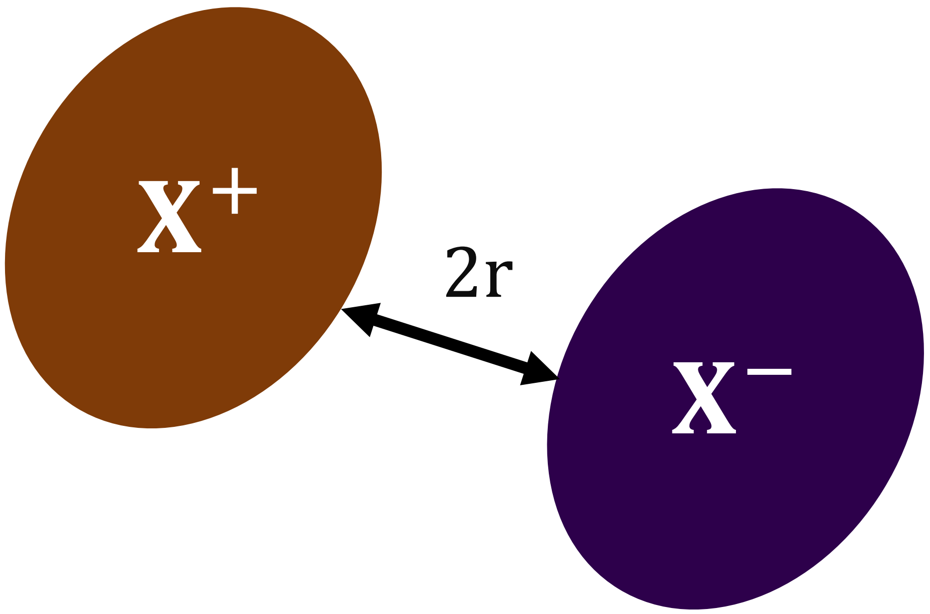
Key idea behind all (provable) results for adversarial robustness

Lemma. If f is L-Locally Lipschitz, then $\mathbf{g} = \mathbf{sign}(f)$ is robust at x whenever $|f(x)| \geq \frac{1}{Lr}$

Hein, Andriushchenko. Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation. NeurIPS 2017.
Cohen, Rosenfeld, Kolter. Certified Adversarial Robustness via Randomized Smoothing. ICML 2019.
Salman, Yang, Li, Zhang, Zhang, Razenshteyn, Bubeck. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. NeurIPS 2019.

Perfect accuracy & robustness, at least in theory

Theorem. If data is $2\mathrm{r}-$ separated, there always exists a classifier that is perfectly robust and accurate, based on a function with local Lipschitz constant $1/r$.



Perfect accuracy & robustness, at least in theory

Theorem. If data is $2r$ – separated, there always exists a classifier that is perfectly robust and accurate, based on a function with local Lipschitz constant $1/r$.

Proof. Classifier: $g = \text{sign}(f)$ where $f(\boldsymbol{x}) = \frac{d(\boldsymbol{x},\mathcal{X}^-)-d(\boldsymbol{x},\mathcal{X}^+)}{2r}$

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') = \frac{d(\boldsymbol{x},\mathcal{X}^-) - d(\boldsymbol{x}',\mathcal{X}^-) - d(\boldsymbol{x},\mathcal{X}^+) + d(\boldsymbol{x}',\mathcal{X}^+)}{2r}$$

$$\leq \frac{2d(\boldsymbol{x},\boldsymbol{x}')}{2r}$$

Lemma → Robust + Accurate if $|f(\boldsymbol{x})| \geq 1$

Robustness is more convoluted...

Adversarial Training: $\min_f \mathbb{E}\left\{ \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X},\epsilon)} \mathcal{L}(f(\mathbf{X}'), Y) \right\}$

Gradient Regularization: $\min_f \mathbb{E}\left\{ \mathcal{L}(f(\mathbf{X}), Y) + \beta \|\nabla_{\mathbf{X}} \mathcal{L}(f(\mathbf{X}), Y)\|_2^2 \right\}$

TRADES: $\min_f \mathbb{E}\left\{ \mathcal{L}(f(\mathbf{X}), Y) + \beta \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X},\epsilon)} \mathcal{L}(f(\mathbf{X}), f(\mathbf{X}')) \right\}$

Madry, Makelov, Schmidt, Tsipras, Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. ICML 2018.
Finlay, Oberman. Scaleable Input Gradient Regularization for Adversarial Robustness. 2019.
Zhang, Yu, Jiao, Xing, Ghaoui, Jordan. Theoretically Principled Trade-off Between Robustness and Accuracy. ICML 2019.

CIFAR-10 Results

methods	train accuracy	test accuracy	adv test accuracy	test lipschitz
Natural	100.00	93.81	0.00	425.71
GR	94.90	80.74	21.32	28.53
LLR	100.00	91.44	22.05	94.68
AT	99.84	83.51	43.51	26.23
TRADES(β=1)	99.76	84.96	43.66	28.01
TRADES(β=3)	99.78	85.55	46.63	22.42
TRADES(β=6)	98.93	84.46	48.58	13.05

Conclusion

New attacks + defense for k-NN, DT, RF

Theory for well-separated data and local Lipschitzness

Q1: Better attack/defense for Random Forests?

Q2: How to achieve high accuracy AND robustness?

Q3: Beyond local Lipschitzness for provable robustness?

Thanks!

Cyrus Rashtchian

www.cyrusrashtchian.com

UCSD

new blog: `ucsdml.github.io`

@CyrusRashtchian

Restricted ImageNet Dataset (1.3M Images)

Classes used in the Restricted ImageNet model. The class ranges are inclusive.

Class	Corresponding ImageNet Classes
"Dog"	151 to 268
"Cat"	281 to 285
"Frog"	30 to 32
"Turtle"	33 to 37
"Bird"	80 to 100
"Primate"	365 to 382
"Fish"	389 to 397
"Crab"	118 to 121
"Insect"	300 to 319

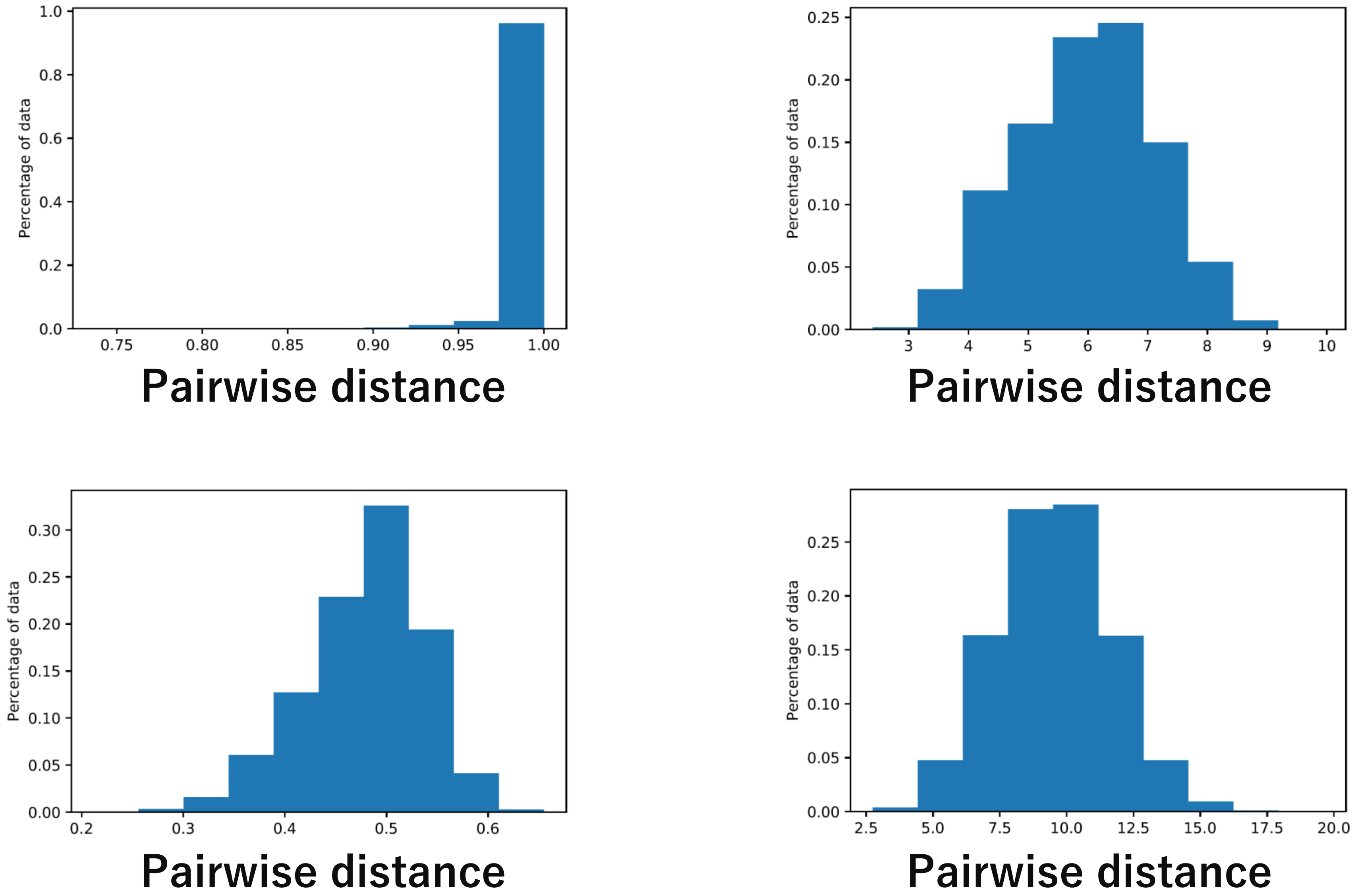
Tsipras, Santurkar, Engstrom,Turner, Madry. Robustness May Be at Odds with Accuracy. ICLR 2019.

Separation of real datasets

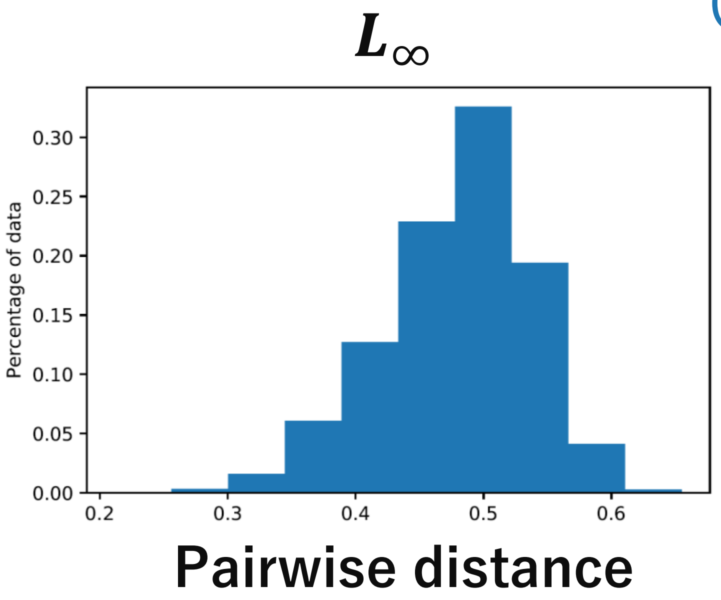
L_∞ L_2

MNIST

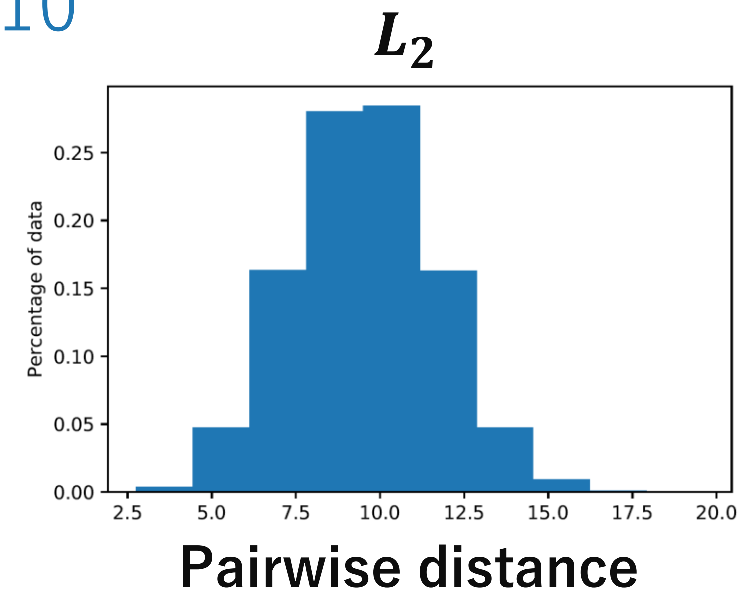
CIFAR-10



CIFAR-10



L_∞
Percentage of data
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.2
0.3
0.4
0.5
0.6
Pairwise distance



L_2
Percentage of data
0.00
0.05
0.10
0.15
0.20
0.25
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Pairwise distance

MNIST Results

methods	train accuracy	test accuracy	adv test accuracy	test lipschitz
Natural	100.00	99.20	59.83	67.25
GR	99.99	99.29	91.03	26.05
LLR	100.00	99.43	92.14	30.44
AT	99.98	99.31	97.21	8.84
TRADES(β=1)	99.81	99.26	96.60	9.69
TRADES(β=3)	99.21	98.96	96.66	7.83
TRADES(β=6)	97.50	97.54	93.68	2.87

ResImageNet Results

Restricted ImageNet	train accuracy	test accuracy	adv test accuracy	test lipschitz
Natural	97.72	93.47	7.89	32228.51
GR	91.12	88.51	62.14	886.75
LLR	98.76	93.44	52.65	4795.66
AT	96.22	90.33	82.25	287.97
TRADES(β=1)	97.39	92.27	79.90	2144.66
TRADES(β=3)	95.74	90.75	82.28	396.67
TRADES(β=6)	93.34	88.92	82.13	200.90

NP Hard for RFs

To find any adversarial example

Reduction from 3SAT

clauses = # trees

(depth 3)

$$(x_0 \vee \neg x_1 \vee x_2) \wedge (x_1 \vee x_3 \vee \neg x_4) \wedge \cdots$$

[Kantchelian, Tyger, Joseph ICML'16]

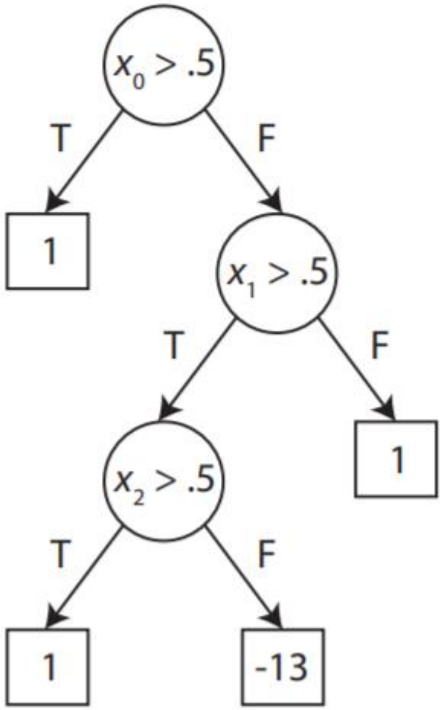
NP Hard for RFs

To find any adversarial example

Reduction from 3SAT

clauses = # trees

(depth 3)



$$(x_0 \vee \neg x_1 \vee x_2)$$

+1 if clause is True
-T if clause is False

$$(x_0 \vee \neg x_1 \vee x_2) \wedge (x_1 \vee x_3 \vee \neg x_4) \wedge \cdots$$

[Kantchelian, Tyger, Joseph ICML'16]

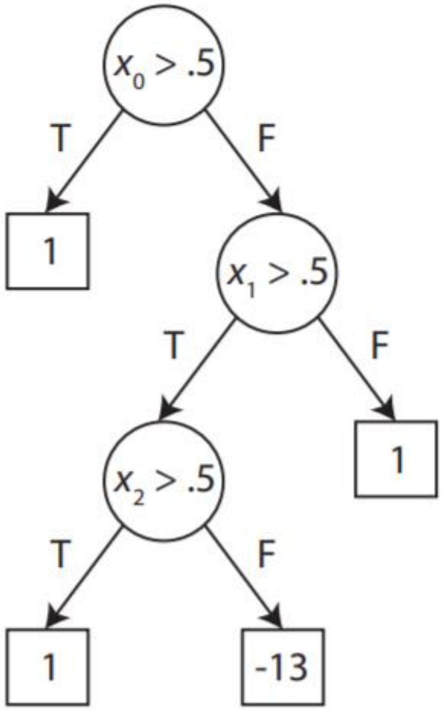
NP Hard for RFs

To find any adversarial example

Reduction from 3SAT

clauses = # trees

(depth 3)



RF outputs True

if and only if

formula is SAT

$$(x_0 \vee \neg x_1 \vee x_2)$$

+1 if clause is True
-T if clause is False

$$(x_0 \vee \neg x_1 \vee x_2) \wedge (x_1 \vee x_3 \vee \neg x_4) \wedge \cdots$$

[Kantchelian, Tyger, Joseph ICML'16]