A Closer Look at Accuracy vs. Robustness

Yao-Yuan Yang

University of California, San Diego

January 25, 2021

joint work with Cyrus Rashtchian, Kamalika Chaudhuri, Ruslan Salakhutdinov and Hongyang Zhang

Introduction

Goodfellow et al. [8]

Lin et al. [10]

Adversarial example

"panda"
57.7% confidence

"nematode"
8.2% confidence

"gibbon"
99.3 % confidence

Goodfellow et al. [9]

gibbon

Adversarial example (cont.)

Not just vision tasks...

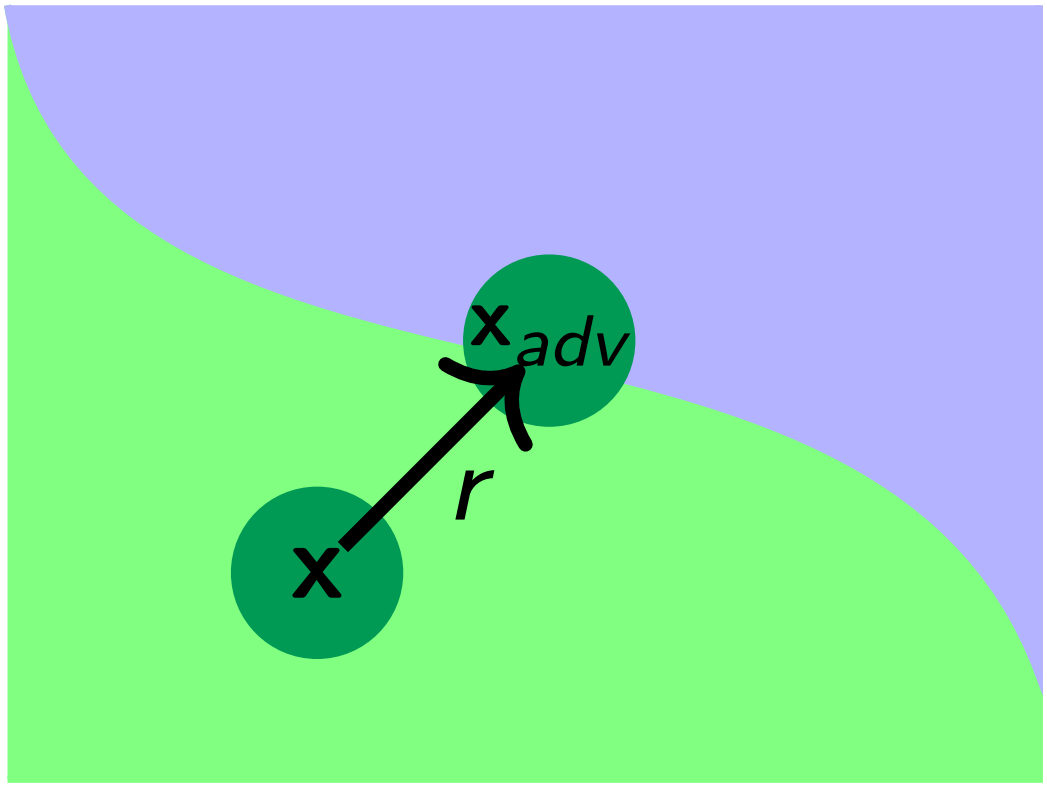
- speech recognition [13]
- natural language processing [18]
- reinforcement learning [1, 2]

Setup

Definition (Adversarial example)

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

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

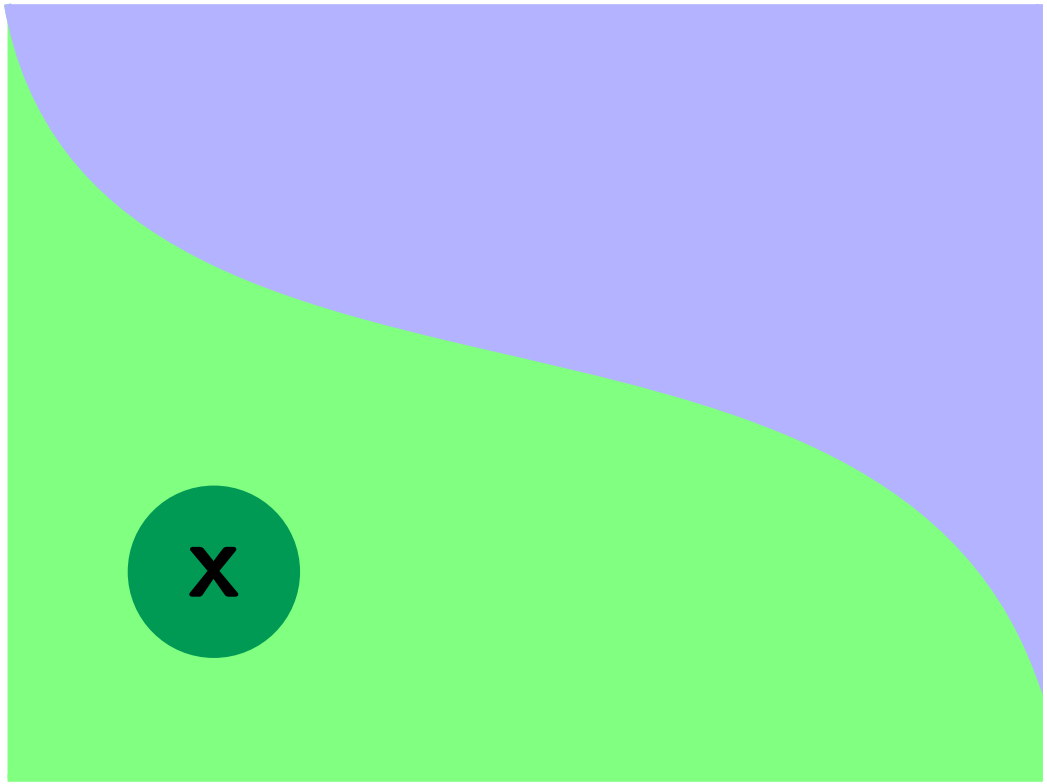


Setup (cont.)

Definition (Accuracy)

The *accuracy* of f under a distribution μ is

$$\Pr_{(\mathbf{x},y)\sim\mu}[f(\mathbf{x}) = y].$$

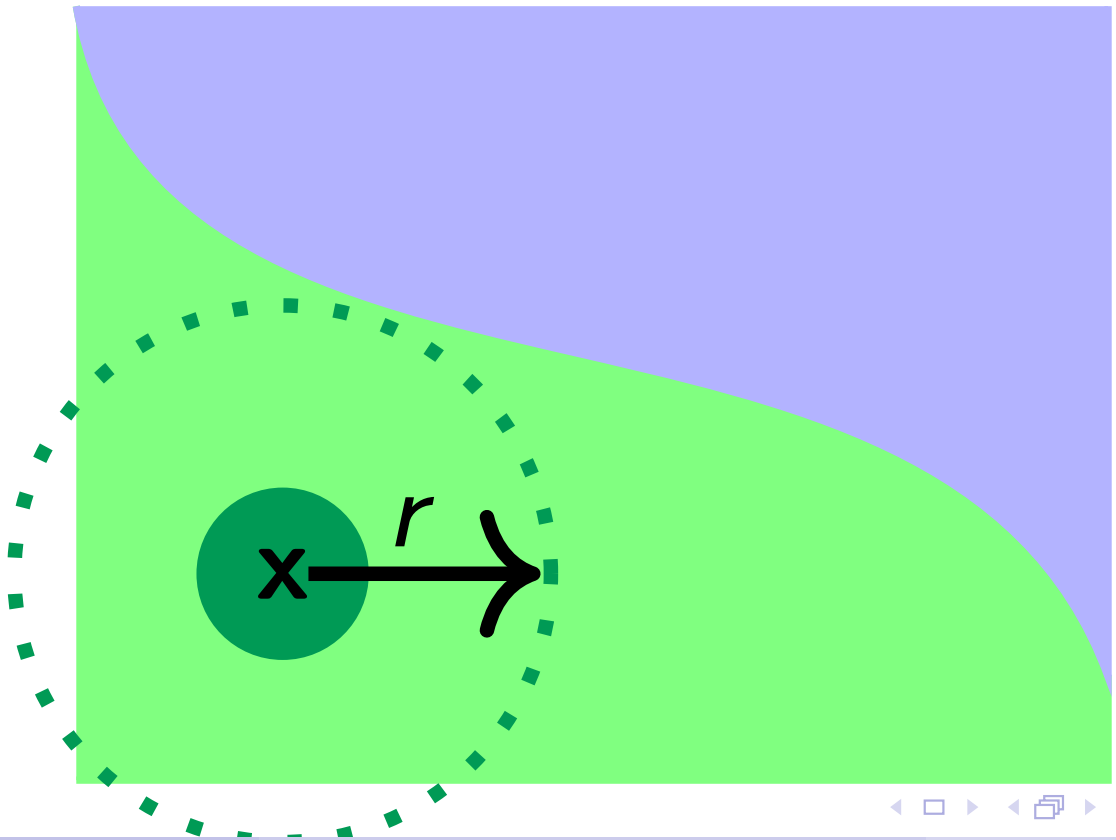


Setup (cont.)

Definition (Astuteness)

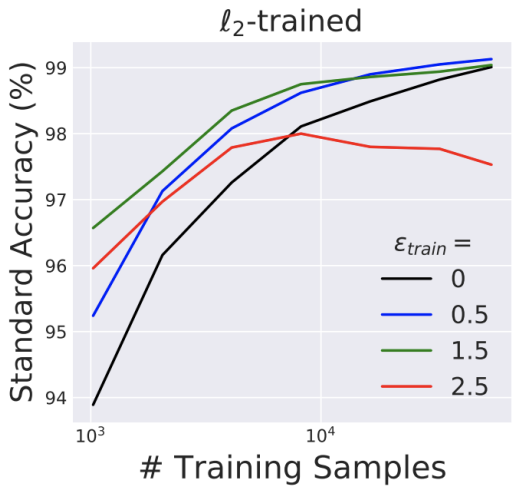
The *astuteness* of f at radius $r > 0$ under a distribution μ is

$$\Pr_{(\mathbf{x},y)\sim\mu}[f(\mathbf{x}') = y \text{ for all } \mathbf{x}' \in \mathbb{B}(\mathbf{x}, r)].$$

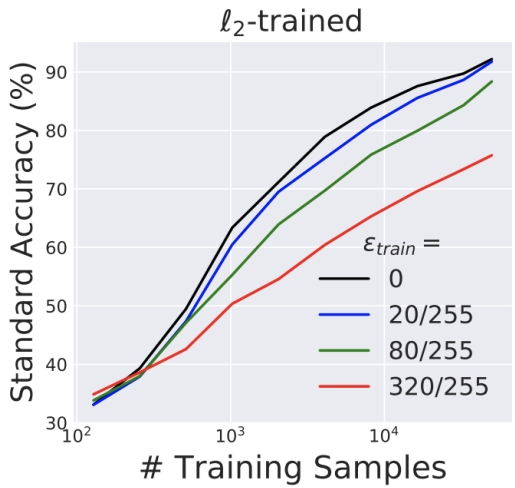


Accuracy robustness trade-off

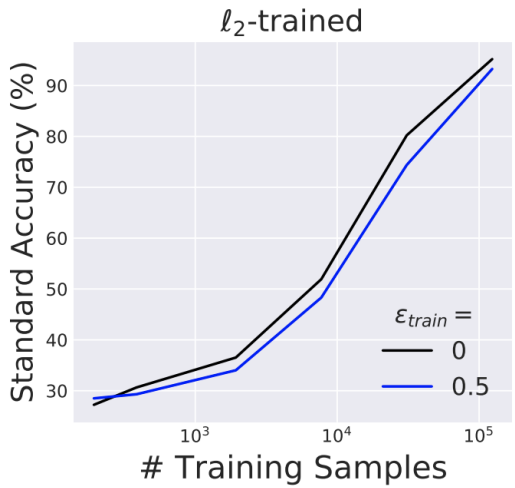
Tsipras et al. [14]



(a) MNIST



(b) CIFAR-10



(c) Restricted ImageNet

and many other works

Zhang et al. [17]
Fawzi et al. [5]
Gilmer et al. [7]

Accuracy robustness trade-off

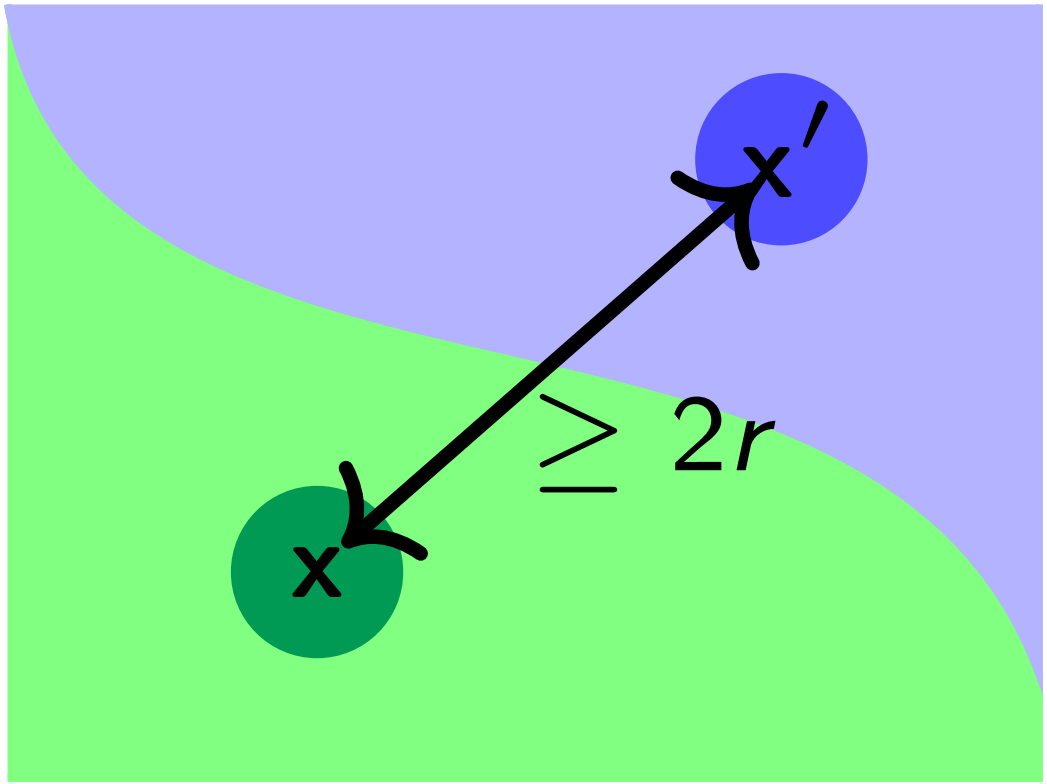
Is this trade-off an intrinsic property?

Accuracy robustness trade-off

Is this trade-off an intrinsic property?

When is this trade-off unavoidable? and when is it not unavoidable?

When differently-labeled examples are separated, we can fit a decision boundary with enough margin



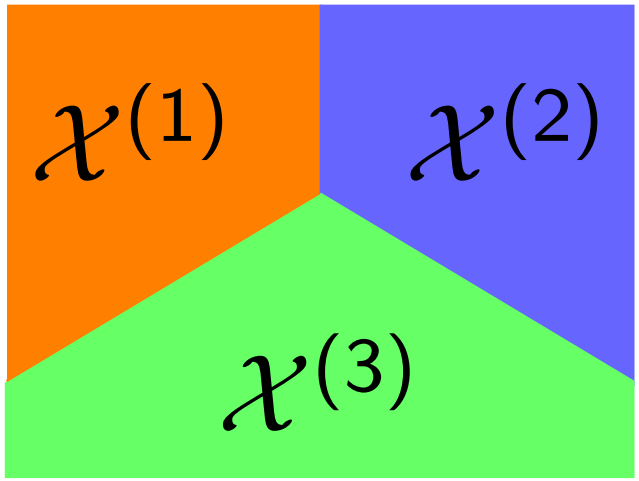
r-separated data

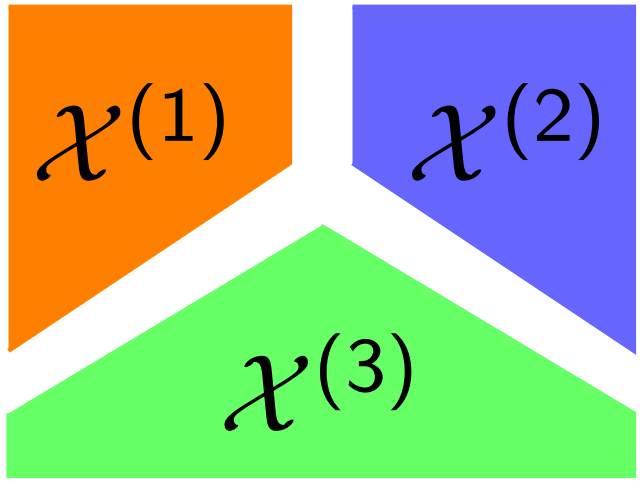
Definition

$\mathcal{X}$ contain C disjoint classes $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(C)}$

Definition (r-separation)

a data distribution over $\bigcup_{i \in [C]} \mathcal{X}^{(i)}$ is r-separated if $\text{dist}(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}) \geq 2r$ for all $i \neq j$, where $\text{dist}(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}) = \min_{\mathbf{x} \in \mathcal{X}^{(i)}, \mathbf{x}' \in \mathcal{X}^{(j)}} \text{dist}(\mathbf{x}, \mathbf{x}')$.





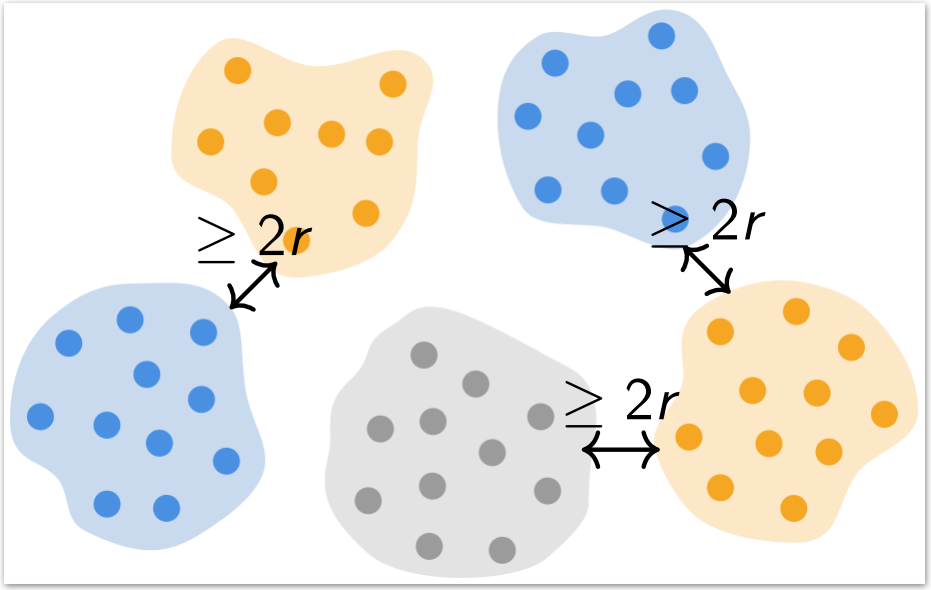
r-separated data

Definition

$\mathcal{X}$ contain C disjoint classes $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(C)}$

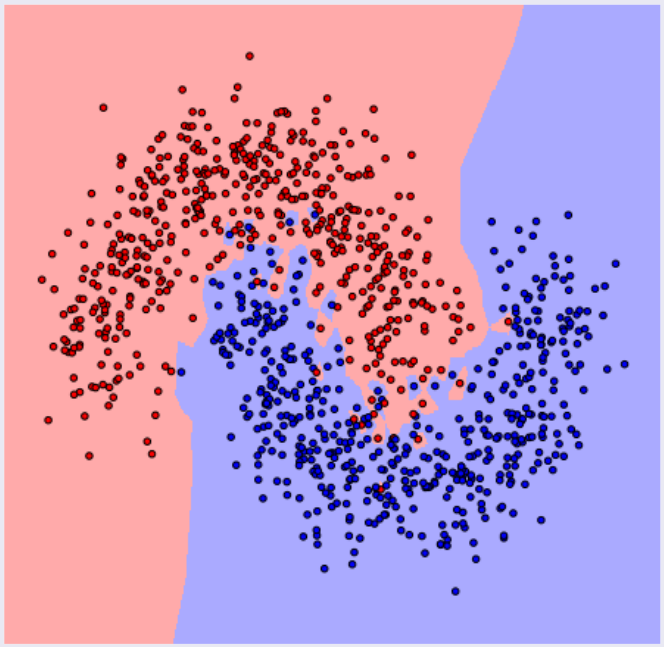
Definition (r-separation)

a data distribution over $\bigcup_{i \in [C]} \mathcal{X}^{(i)}$ is r-separated if $\text{dist}(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}) \geq 2r$ for all $i \neq j$, where $\text{dist}(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}) = \min_{\mathbf{x} \in \mathcal{X}^{(i)}, \mathbf{x}' \in \mathcal{X}^{(j)}} \text{dist}(\mathbf{x}, \mathbf{x}')$.

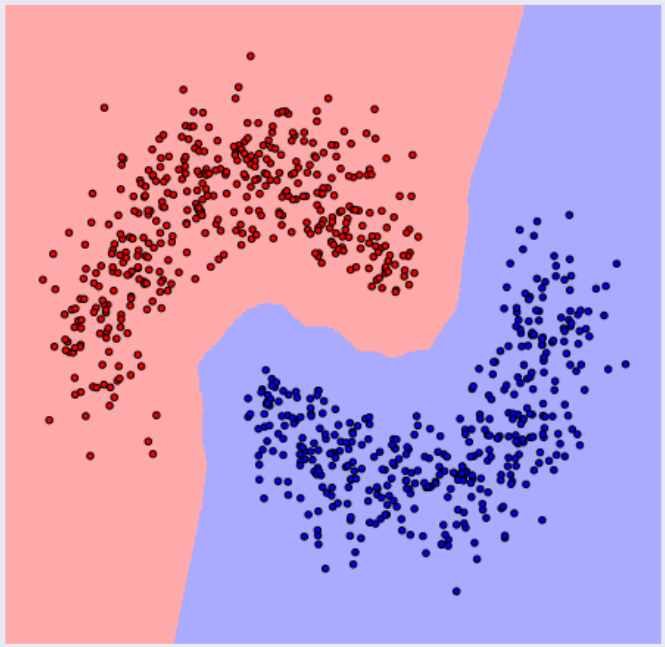


Types of datasets

Not r-separated

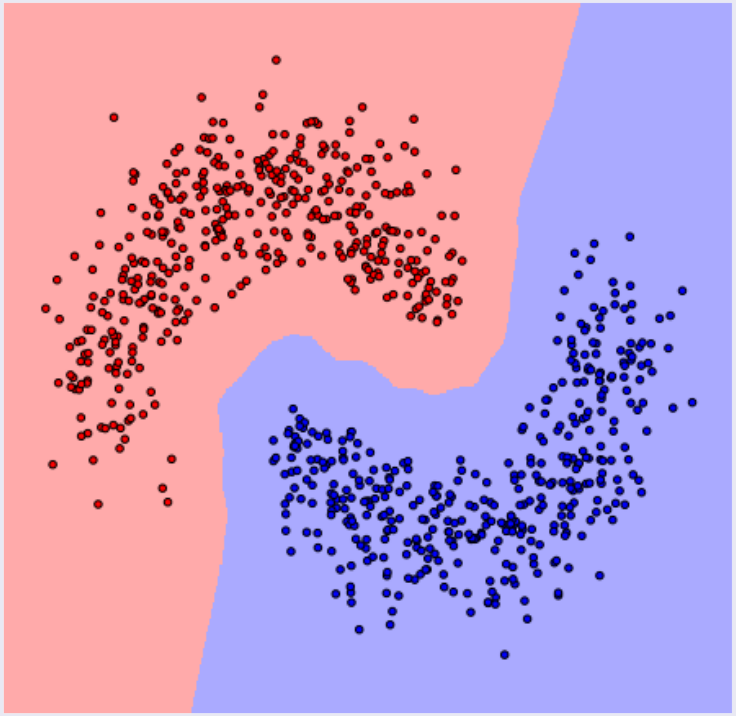


r-separated



r-separated data

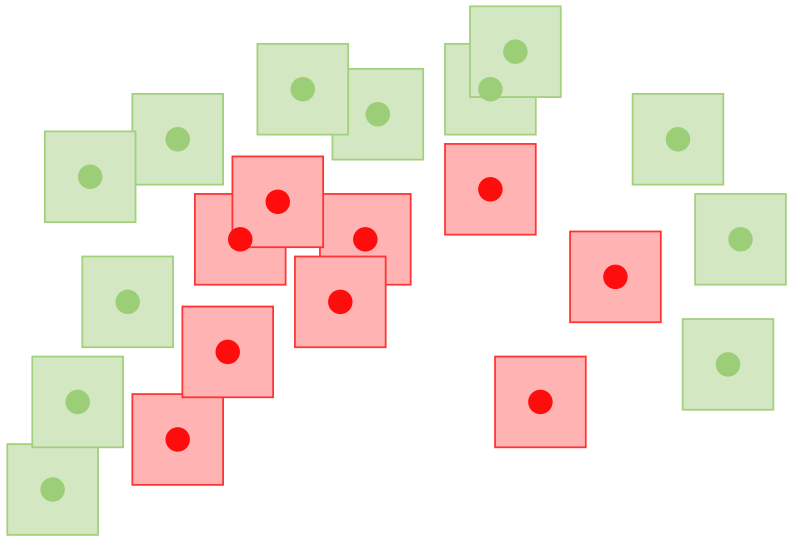
For separated data, it is possible to fit a boundary between differently labeled examples



No intrinsic trade-off for r-separated data

r-separated data implies the existance a classifier that is

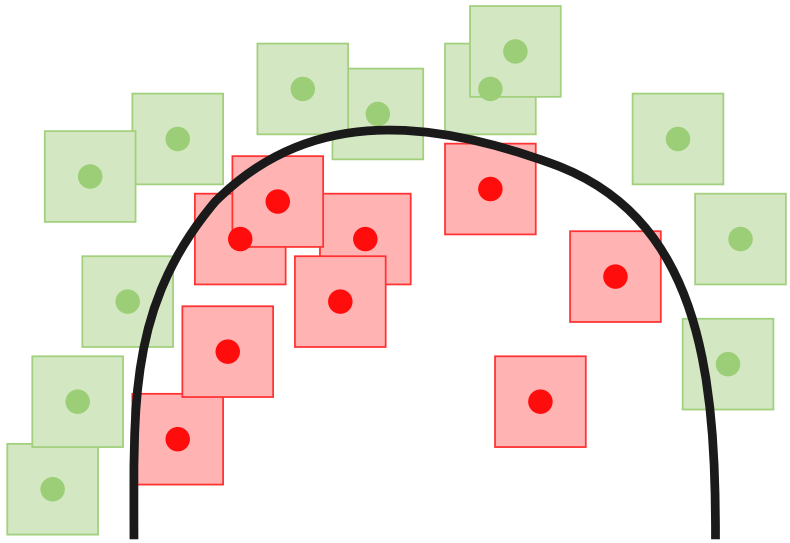
- Accurate
- Robust
- Locally smooth (locally Lipschitz)



No intrinsic trade-off for r-separated data

r-separated data implies the existance a classifier that is

- Accurate
- Robust
- Locally smooth (locally Lipschitz)



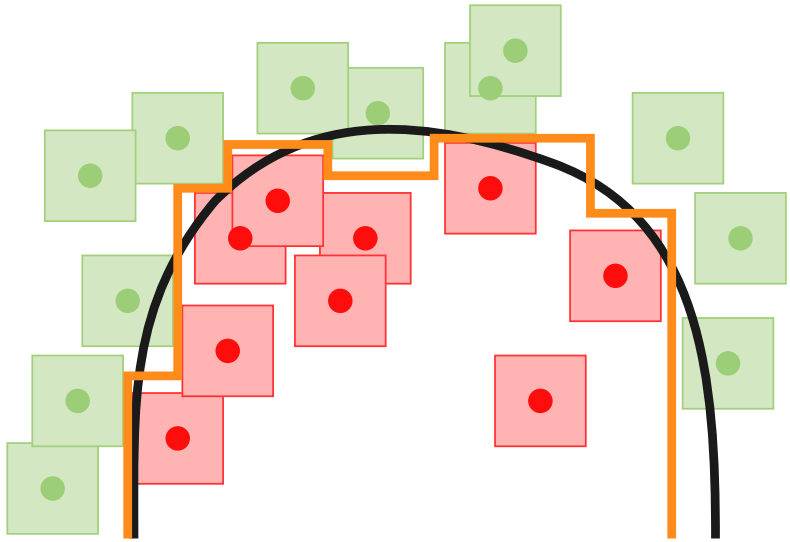
No intrinsic trade-off for r-separated data (cont.)

Definition (L-locally Lipschitz)

A function $f : \mathcal{X} \to \mathbb{R}^C$ is L-locally Lipschitz at $\mathbf{x}$ with radius r if for each $i \in [C]$, we have

$$|f(\mathbf{x})_i - f(\mathbf{x}')_i| \leq L \cdot \text{dist}(\mathbf{x}, \mathbf{x}')$$

for all $\mathbf{x}'$ with $\text{dist}(\mathbf{x}, \mathbf{x}') \leq r$.



No intrinsic trade-off for r-separated data (cont.)

Theorem

If data is r-separated, there always exists a classifier that is perfectly robust and accurate

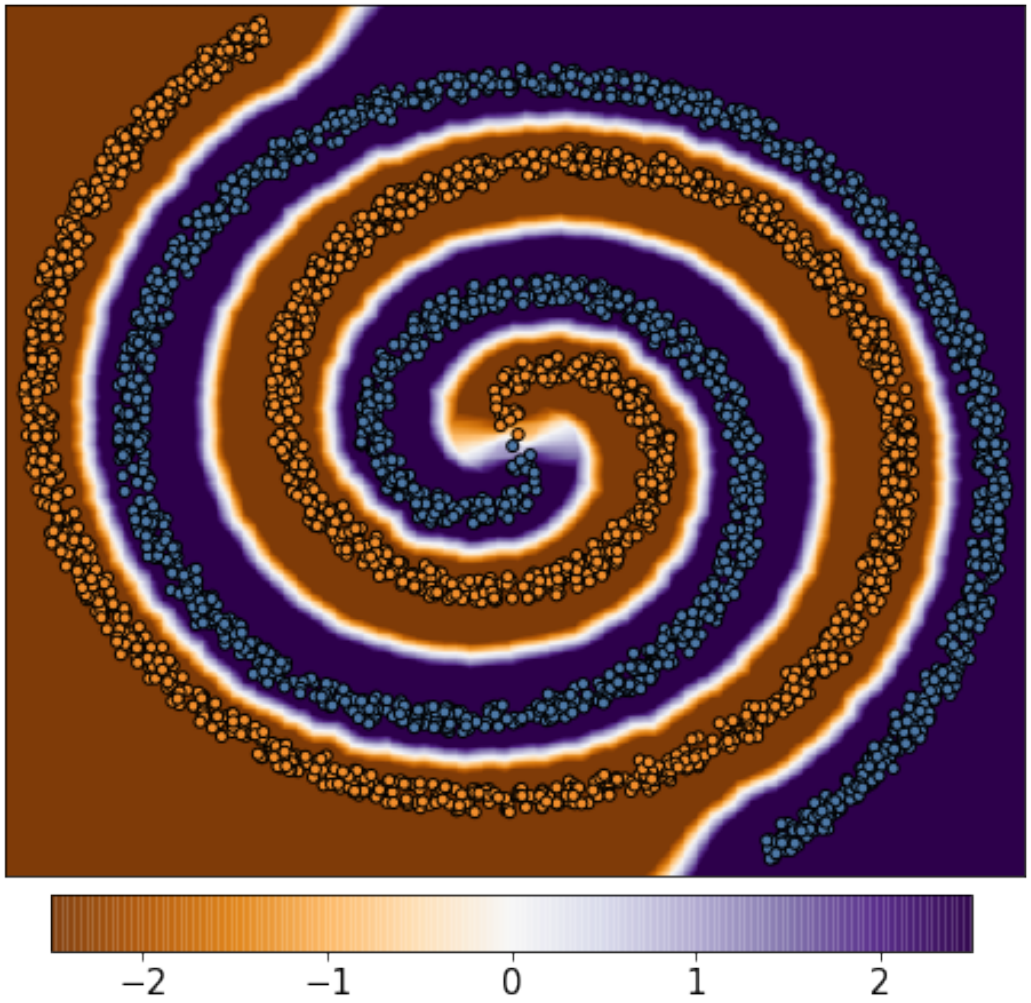
Key idea

Consider $g(\mathbf{x}) = \text{argmin}_{i \in [C]} f(\mathbf{x})_i$, where $f(\mathbf{x})_i = \frac{1}{r}\text{dist}(\mathbf{x}, \mathcal{X}^{(i)})$.

Then we show that g is robust and accurate through local Lipschitzness.

Locally Lipschitz classifier

$$g(\mathbf{x}) = \text{argmin}_{i \in [C]} \frac{1}{r} \text{dist}(\mathbf{x}, \mathcal{X}^{(i)})$$

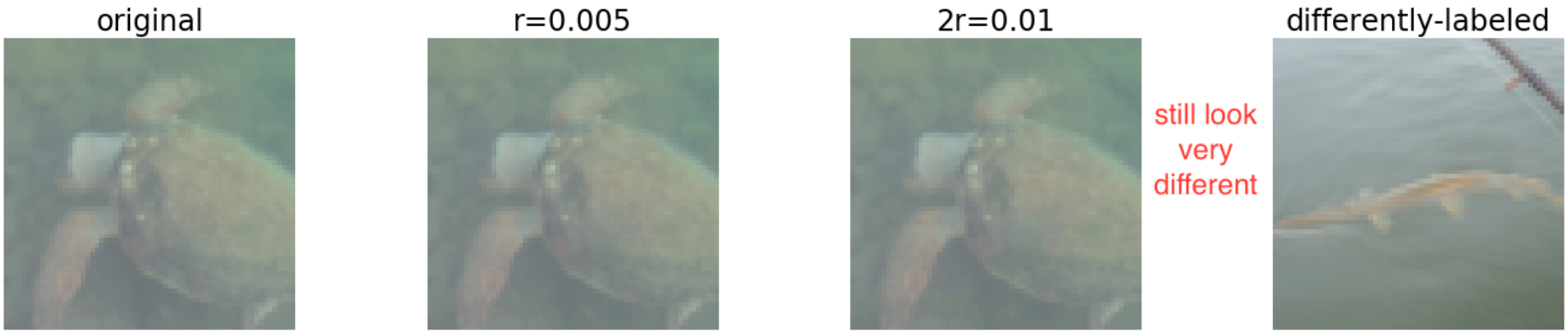


r-separation implies the existence of an astute classifier

r-separation implies the existence of an astute classifier

Are real datasets r-separated?

Separation for image datasets



Separation for image datasets (cont.)

Many benchmark datasets are separated enough for common robust radius (r) to be both robust and accurate

	r	Required separation ($2r$)	min Train-Train separation	min Test-Train separation
MNIST	0.1	0.2	0.737	0.812
CIFAR-10	0.031 (8/255)	0.062	0.212	0.220
SVHN	0.031 (8/255)	0.062	0.094	0.110
ResImageNet	0.005	0.01	0.180	0.224

Many image datasets are separated, thus astute classifiers exist for them

Many image datasets are separated, thus astute classifiers exist for them

- Recall that previous works have conjectured robustness accuracy trade-off being inevitable
- It appears that there is a gap between theory and practice
- What might be the cause this gap?

Many image datasets are separated, thus astute classifiers exist for them

- Recall that previous works have conjectured robustness accuracy trade-off being inevitable
- It appears that there is a gap between theory and practice
- What might be the cause this gap?

- Is local Lipschitzness correlated with robustness?

A closer look at existing defense methods

low local lipschitz (smooth) classifier:

- generates higher adversarial test accuracy
- increases the generalization gap

		CIFAR-10					
		train acc.	test acc.	adv test acc.	test lipschitz	gap	adv gap
high lip.	natural	100.00	93.81	0.00	425.71	6.19	0.00
	GR [6]	94.90	80.74	21.32	28.53	14.16	3.94
	LLR [12]	100.00	91.44	22.05	94.68	8.56	4.50
low lip.	RST [16]	99.86	84.61	40.89	23.15	15.25	41.31
	AT [11]	99.84	83.51	43.51	26.23	16.33	49.94
	TRADES [17]	99.78	85.55	46.63	22.42	14.23	47.67

Improving generalization with dropout

Standard tools like dropout can improve generalization. However, it may not be enough.

		CIFAR-10				
	dropout	test acc.	adv test acc.	test lipschitz	gap	adv gap
Natural	w/o dropout	93.81	0.00	425.71	6.19	0.00
Natural	w/ dropout	93.87	0.00	384.48	6.13	0.00
AT	w/o dropout	83.51	43.51	26.23	16.33	49.94
AT	w/ dropout	85.20	43.07	31.59	14.51	44.05
RST	w/o dropout	83.87	41.75	23.80	15.86	43.54
RST	w/ dropout	85.49	40.24	34.45	14.00	33.07
TRADES	w/o dropout	84.46	48.58	13.05	14.47	42.65
TRADES	w/ dropout	84.69	52.32	8.13	11.91	26.49

Current algorithms do not find the perfectly robust and accurate classifier, possibly because they do not generalize well enough

Need an algorithm that generalizes better

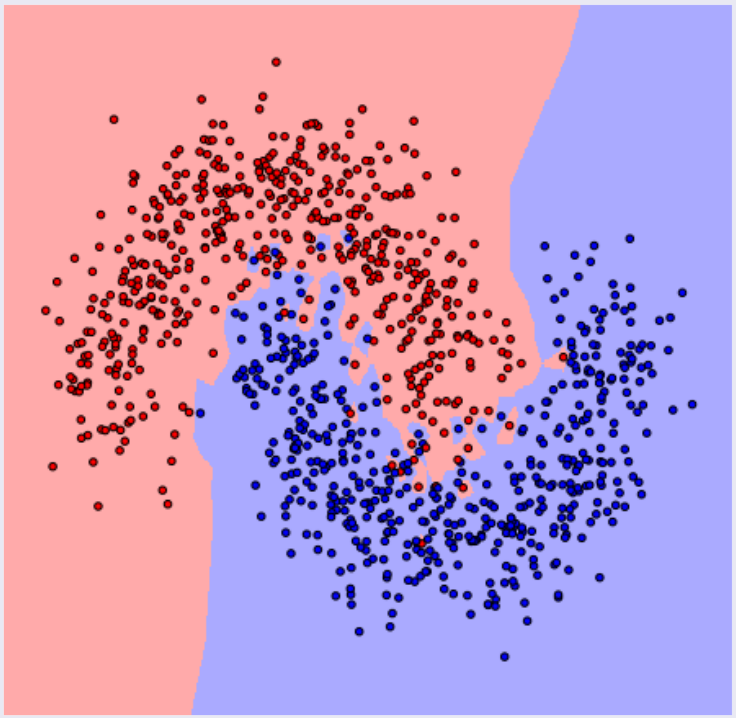
Current algorithms do not find the perfectly robust and accurate classifier, possibly because they do not generalize well enough

Need an algorithm that generalizes better

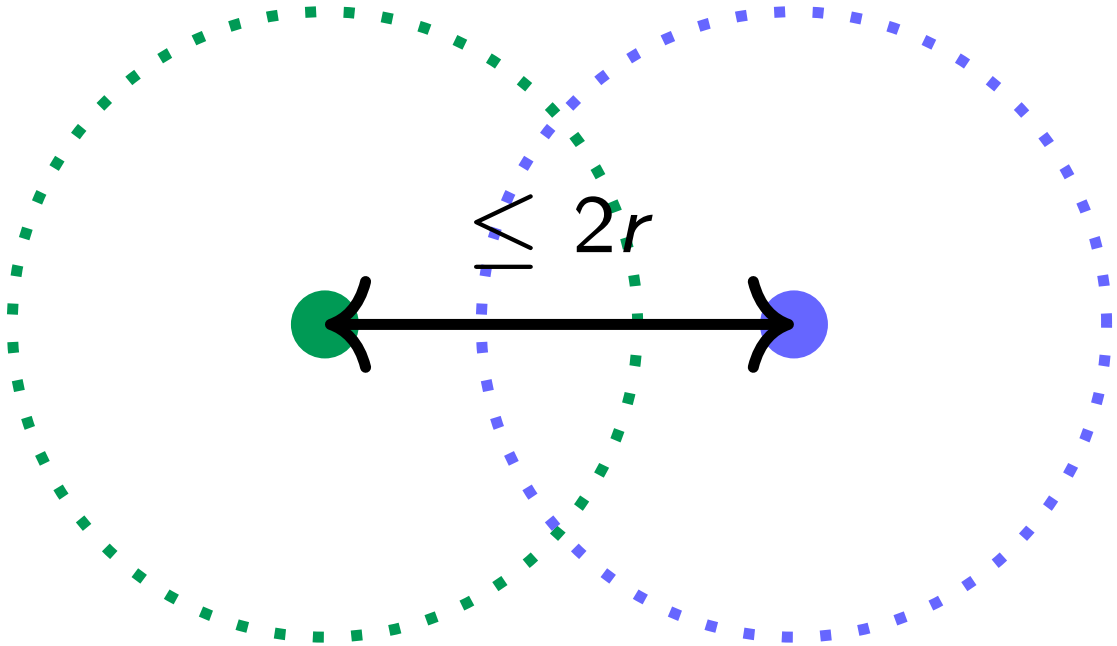
How about datasets that are not r-separated?

Data that are not separated

When data is not r-separated, there maybe no perfectly robust and accurate classifier



What if the data is not separated?



The norm balls are overlapping, so no classifier can predict these two examples both accurately and robustly

Datasets that are not separated

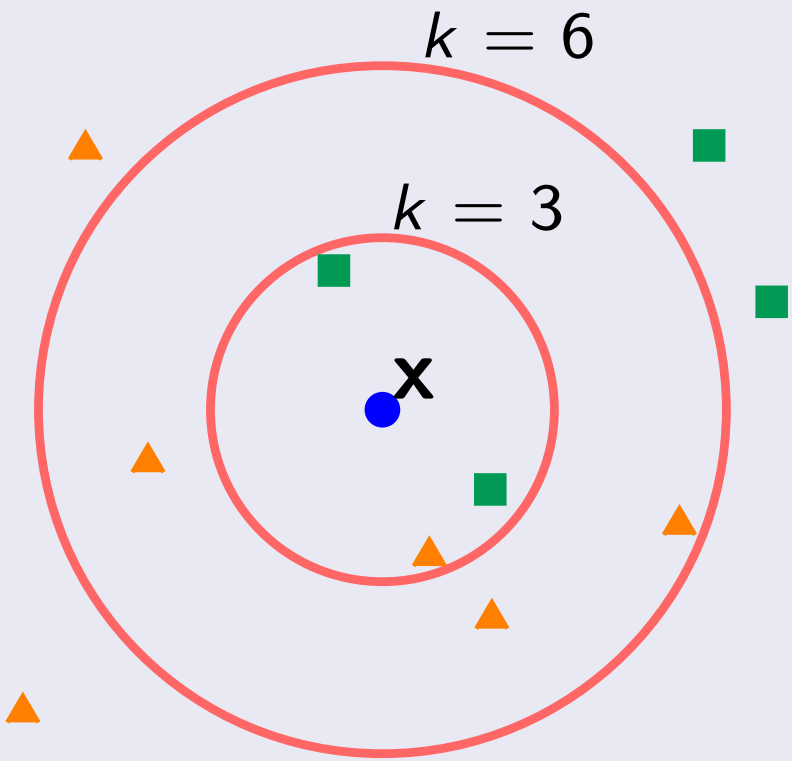
Tabular datasets

Traditional non-parametric classifiers such as k-NN or DT works well on such datasets

Non-parametric classifiers

k nearest neighbor (k-NN)

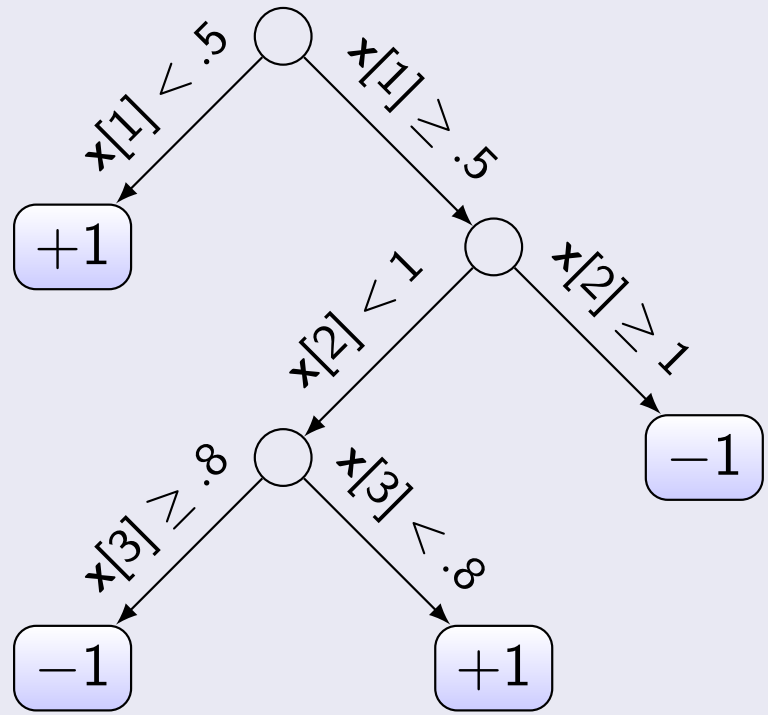
take k closest training examples and output the majority label



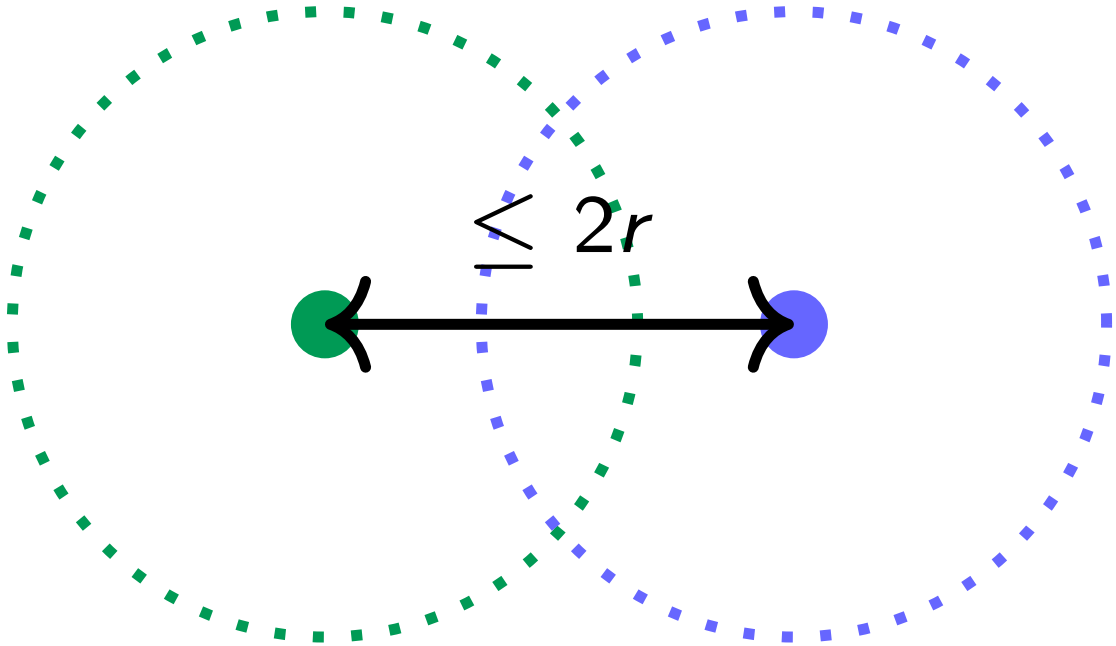
Decision tree and tree ensembles

recursively split the data

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

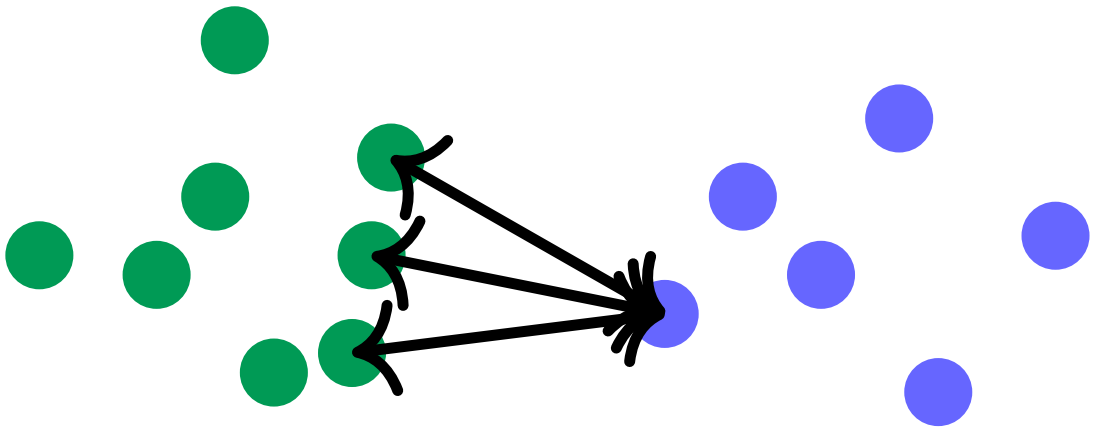


What if the data is not separated?



When requiring robust radius r, what is the highest accuracy we can get?

Adversarial pruning

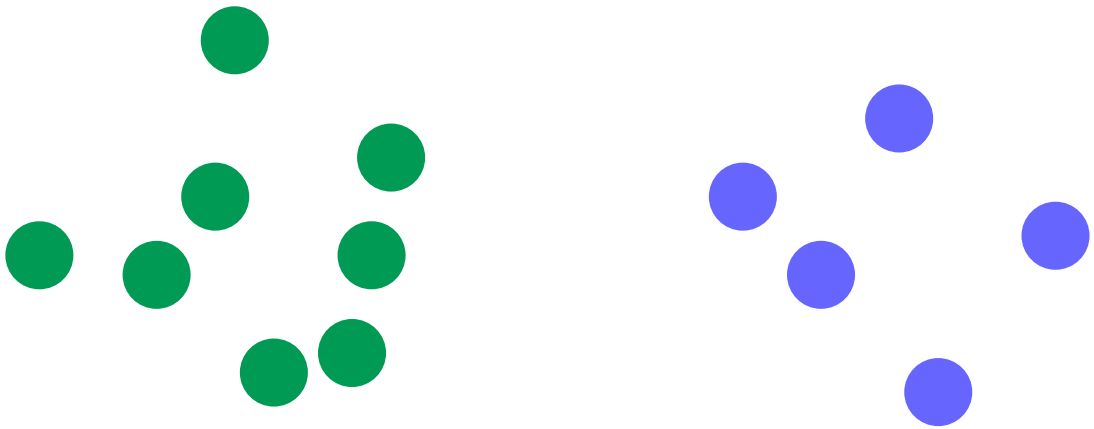


- Under astuteness, a non-robust example is considered as incorrectly predicted
- Remove examples until the data is separated

Adversarial pruning

remove minimum # of examples s.t. distance between differently-labeled examples are $\geq 2r$ (minimum vertex cover problem)

Adversarial pruning (cont.)



- Under astuteness, a non-robust example is considered as incorrectly predicted
- Remove examples until the data is separated

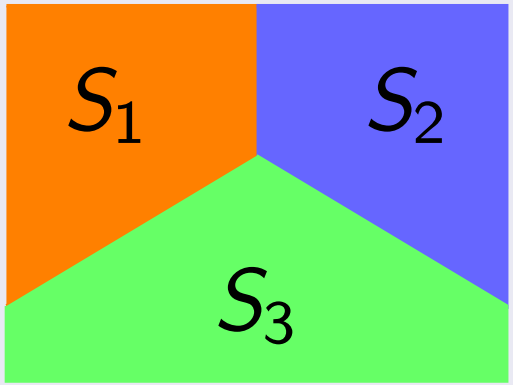
Adversarial pruning

remove minimum # of examples s.t. distance between differently-labeled examples are $\geq 2r$ (minimum vertex cover problem)

Implications for large sample limit

Bayes-optimal classifier

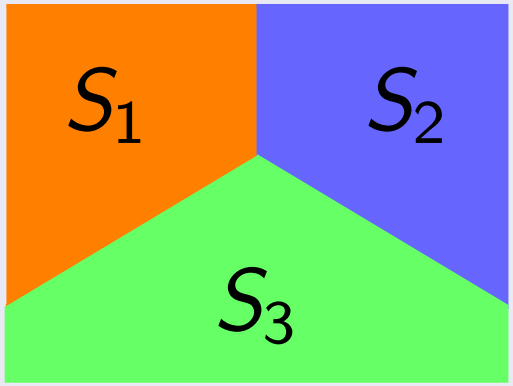
$$\max_{S_1,\ldots,S_c} \sum_{j=1}^{c} \int_{\mathbf{x}\in S_j} Pr(y = j \mid \mathbf{x}) d\mu$$



Implications for large sample limit

Bayes-optimal classifier

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

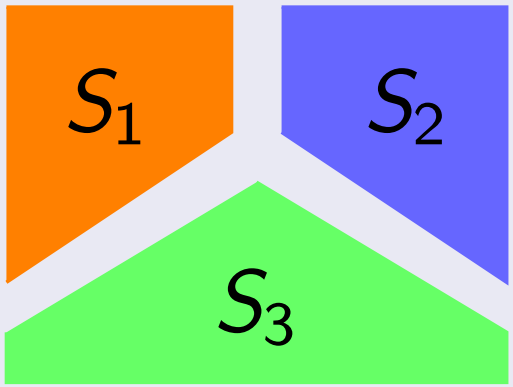


r-Optimal classifier

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

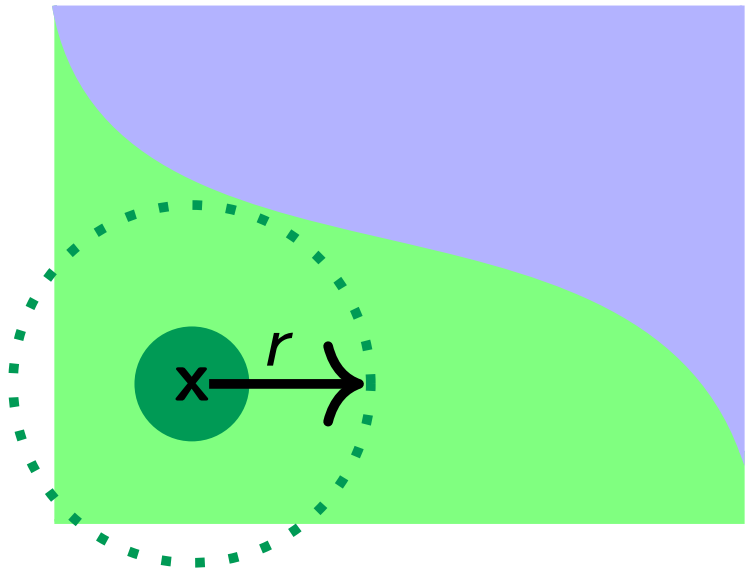
$$\text{s.t.} \quad d(S_j, S_{j'}) \geq 2r \quad \forall j \neq j'$$

$$d(S_j, S_{j'}) := \min_{u\in S_j, v\in S_{j'}} \|u - v\|_p$$



Definition (Astuteness)

The *astuteness* of f at radius $r > 0$ under a distribution μ is $\Pr_{(\mathbf{x},y)\sim\mu}[f(\mathbf{x}_{adv}) = y \text{ for all } \mathbf{x}_{adv} \in \mathbb{B}(\mathbf{x}, r)]$.



Theorem (r-Optimal classifier has the optimal astuteness)

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

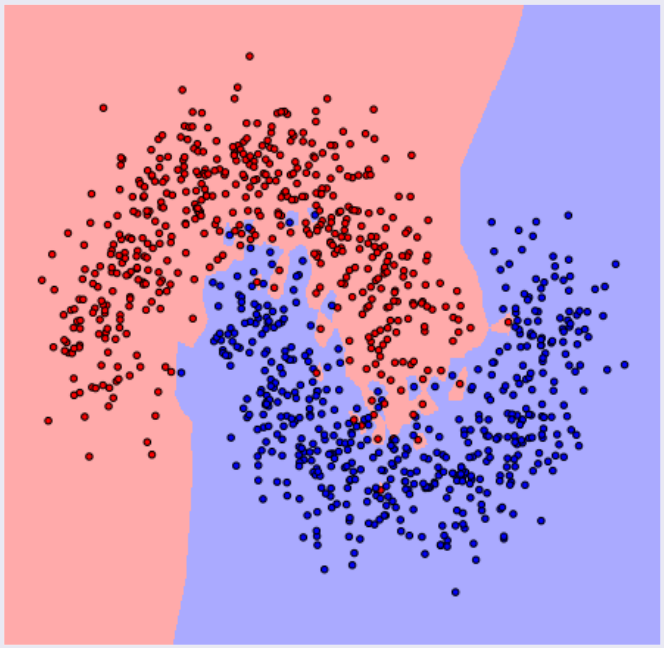
$$f_{\text{ropt}} = \text{argmax}_f \ \text{ast}_\mu(f, r)$$

Follow up theory by Bhattacharjee and Chaudhuri [3]

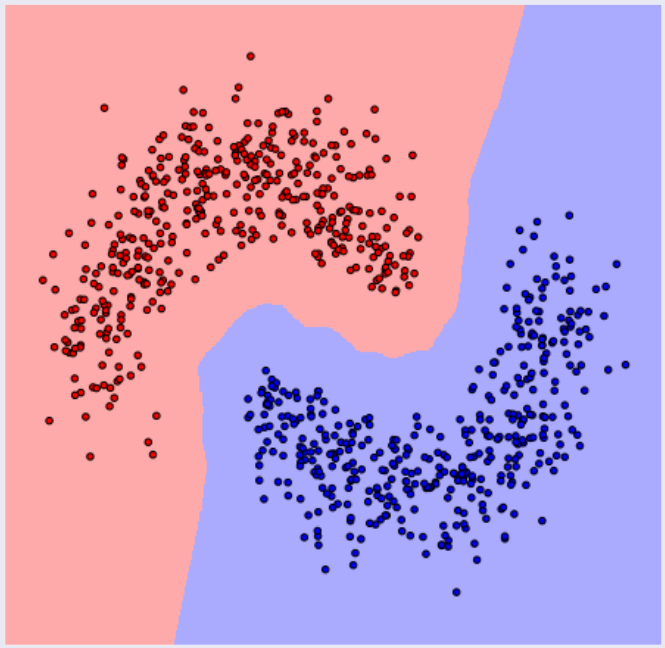
They prove that AP + k-NN or kernel classifiers converges toward optimally robust and accurate classifiers.

Adversarial pruning

Before AP

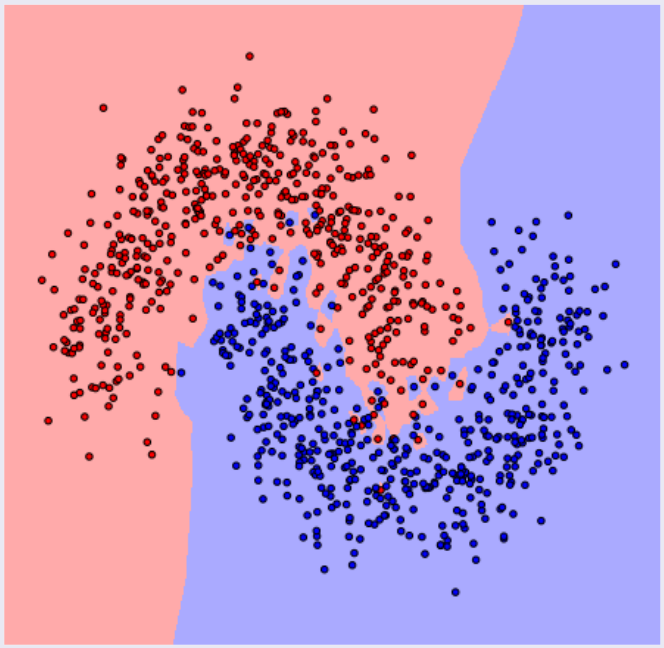


After AP

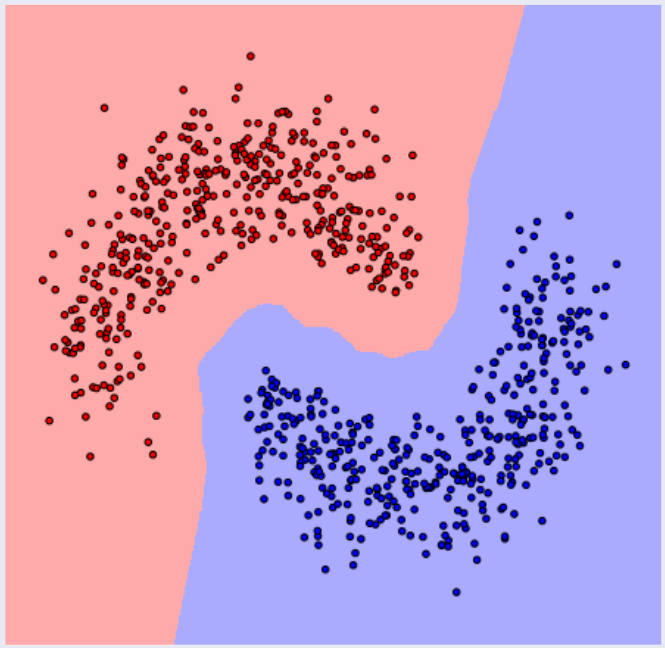


Adversarial pruning

Before AP

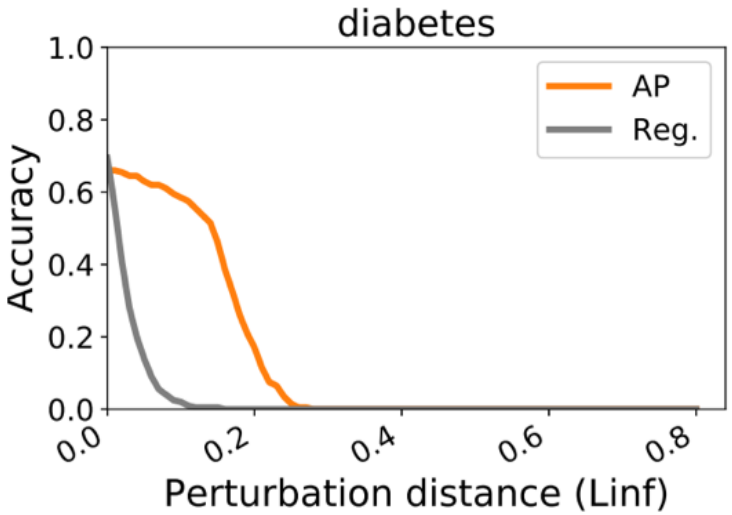


After AP

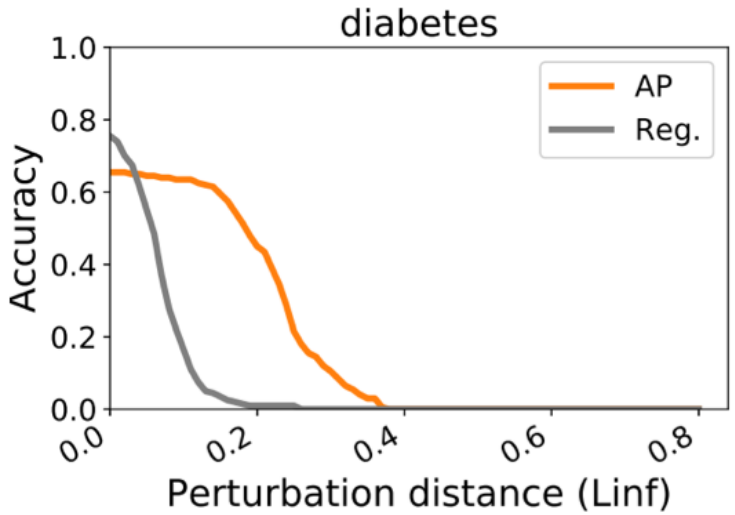


How about on real data?

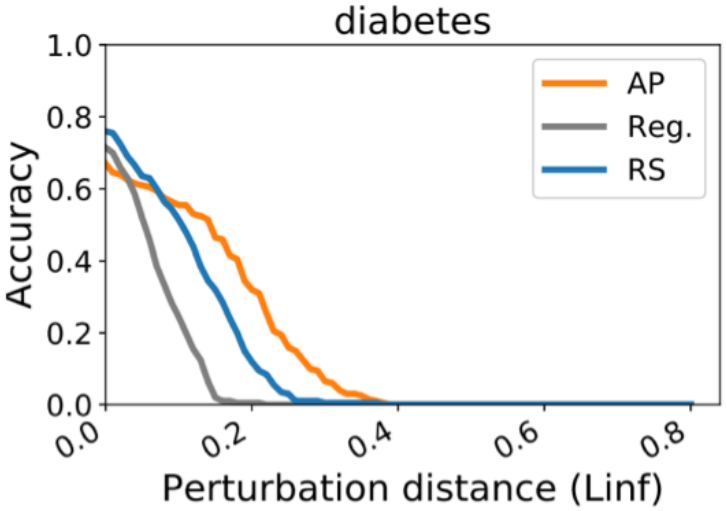
Defense with AP



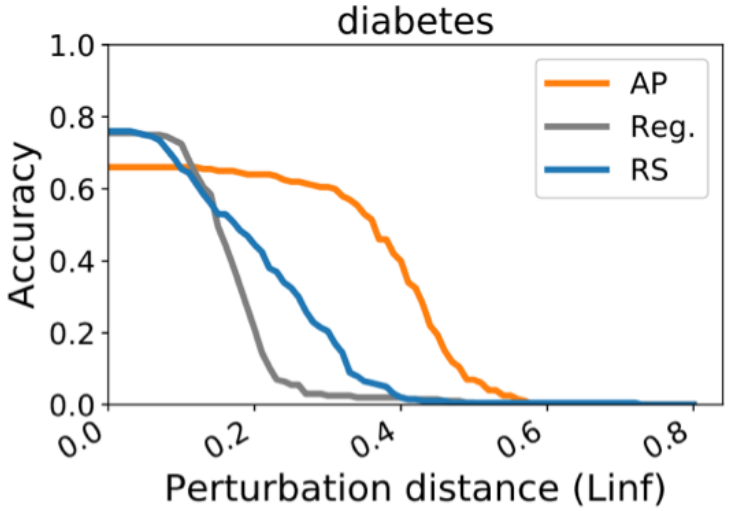
(e) 1-NN



(f) 3-NN



(g) Decision tree



(h) Random forest

Examples removed by AP

	# examples			
	original	AP($r = 0.1$)	AP($r = 0.3$)	AP($r = 0.5$)
australian	490	484	458	427
cancer	483	483	473	458
covtype	2000	1904	1417	1384
diabetes	568	535	379	370
fourclass	662	559	453	442

Conclusion

For many benchmark image classification datasets, it is theoretically possible to achieve perfect robustness and accuracy

Conclusion

For many benchmark image classification datasets, it is theoretically possible to achieve perfect robustness and accuracy

The observed trade-off happens plausibly due to poor generalization

Conclusion

For many benchmark image classification datasets, it is theoretically possible to achieve perfect robustness and accuracy

The observed trade-off happens plausibly due to poor generalization

For non-r-separated data, using adversarial pruning to make the data more separated improves robustness.

References

A Closer Look at Accuracy vs. Robustness. Yao-Yuan Yang*, Cyrus Rashtchian*, Hongyang Zhang, Ruslan Salakhutdinov, Kamalika Chaudhuri, in NeurIPS, 2020

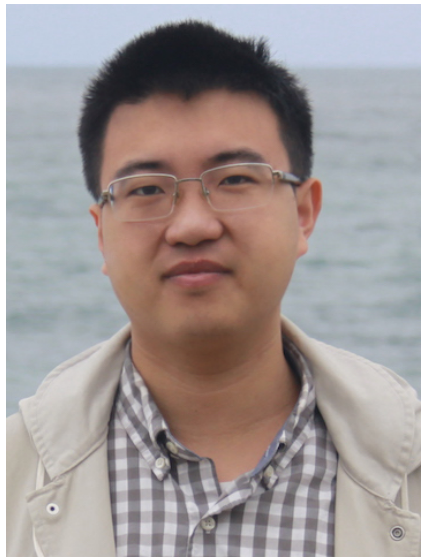
Robustness for Non-Parametric Classification: A Generic Attack and Defense. Yao-Yuan Yang*, Cyrus Rashtchian*, Yizhen Wang, and Kamalika Chaudhuri, in AISTATS, 2020

* equal contribution

Thanks

Kamalika Chaudhuri

Cyrus Rashtchian



Hongyang Zhang

Ruslan Salakhutdinov

Thank you for listening.

More information

- Paper:
  - A Closer Look at Accuracy vs. Robustness: `https://arxiv.org/abs/2003.02460`
  - Robustness for Non-Parametric Classification: A Generic Attack and Defense: `https://arxiv.org/abs/1906.03310`
- Blog: `https://ucsdml.github.io/`

Contact

- Website: `http://yyyang.me/`

References I

[1] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer, 2017.

[2] Vahid Behzadan and Arslan Munir. Whatever does not kill deep reinforcement learning, makes it stronger. *arXiv preprint arXiv:1712.09344*, 2017.

[3] Robi Bhattacharjee and Kamalika Chaudhuri. When are non-parametric methods robust? *arXiv preprint arXiv:2003.06121*, 2020.

[4] Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. Robust Decision Trees Against Adversarial Examples. In *ICML*, 2019.

[5] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *Advances in neural information processing systems*, 31:1178–1187, 2018.

References II

[6] Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

[7] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[8] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

References III

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[12] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pages 13847–13856, 2019.

[13] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. *arXiv preprint arXiv:1903.10346*, 2019.

References IV

[14] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[15] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. In *ICML*, pages 5120–5129, 2018.

[16] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

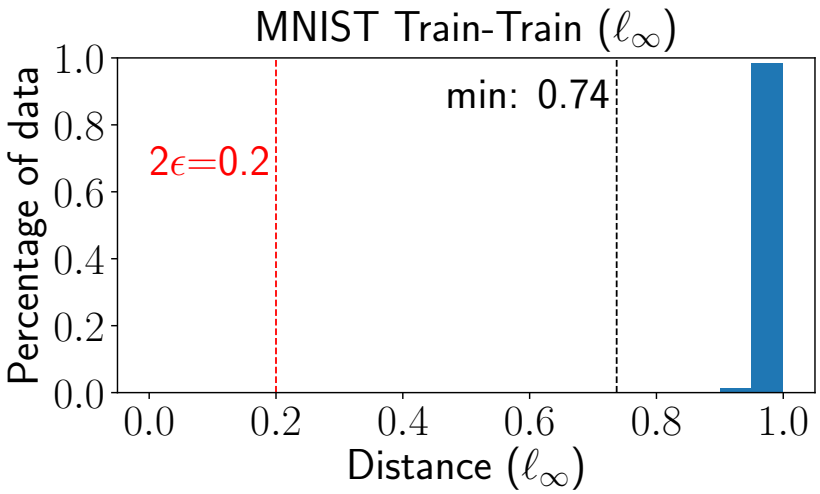
[17] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

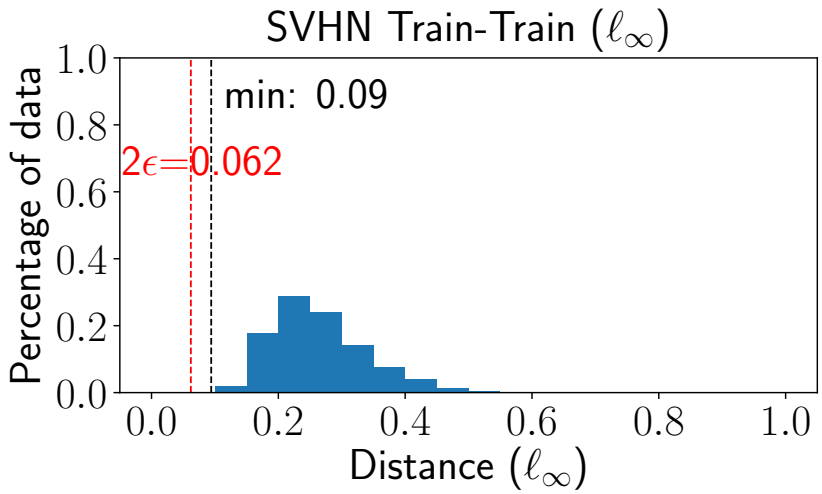
References V

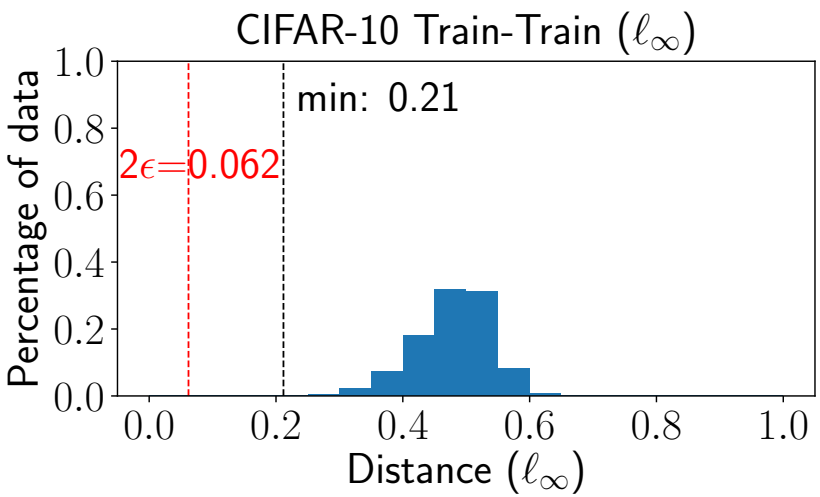
[18] Wei Emma Zhang, Quan Z Sheng, and Ahoud Abdulrahmn F Alhazmi. Generating textual adversarial examples for deep learning models: A survey. *arXiv preprint arXiv:1901.06796*, 2019.

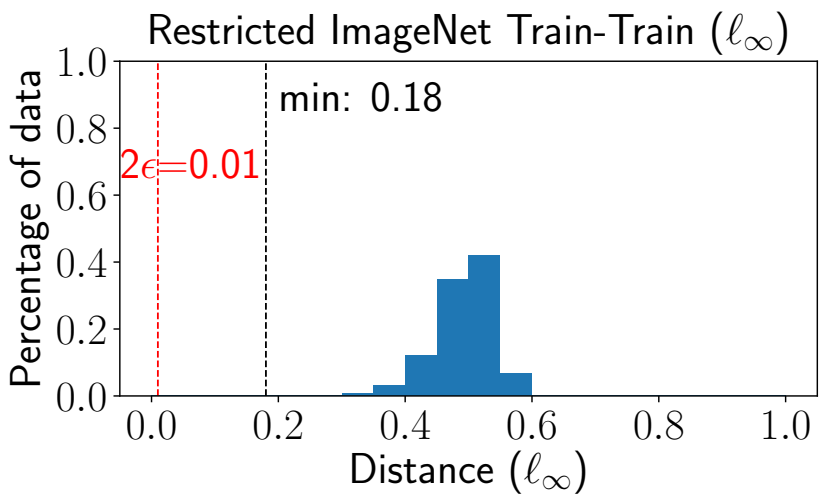
Separation for image datasets

The distribution of $\min_{j:y_j \neq y_i} dist(\mathbf{x}_i, \mathbf{x}_j)$ for each $\mathbf{x}_i$









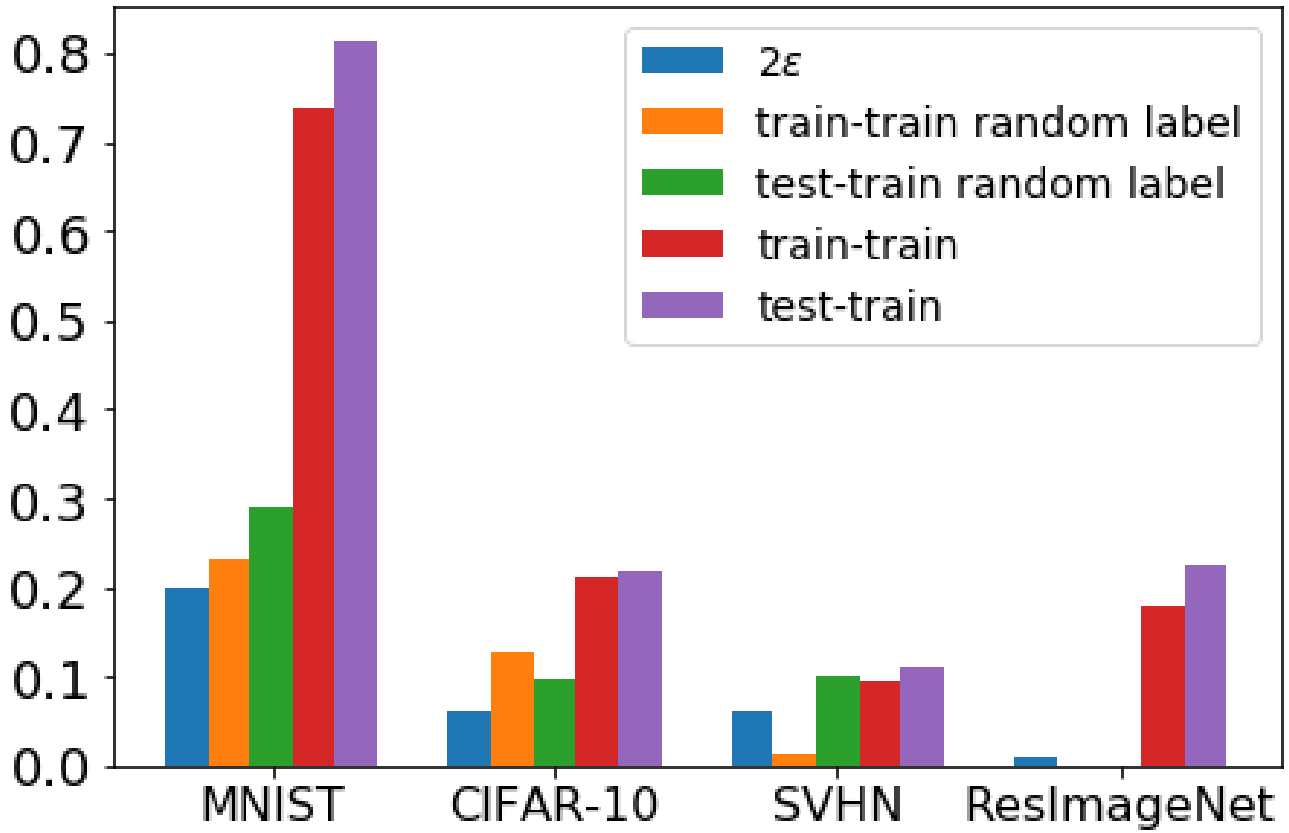
Separation of randomly labeled datasets

	ϵ	randomly labeled				original labels			
		train-train		test-train		train-train		test-train	
		min	mean	min	mean	min	mean	min	mean
MNIST	0.100	0.231	0.902	0.290	0.904	0.737	0.990	0.812	0.990
CIFAR-10	0.031	0.125	0.476	0.098	0.475	0.212	0.479	0.220	0.479
SVHN	0.031	0.012	0.259	0.102	0.271	0.094	0.264	0.110	0.274
ResImageNet	0.005	0.000	0.485	0.000	0.483	0.180	0.492	0.224	0.492

Table: Separation results on real datasets for both original labels and randomly assigned labels.

Separation of randomly-labeled datasets

The separation is much larger between different classes than within the same class



Proof-of-concept classifier

	perturbation ϵ	accuracy	adversarial accuracy
MNIST	0.1	99.99	99.98
SVHN	0.031	100.00	99.90
CIFAR-10	0.031	100.00	99.99

Table: Train with both training and testing data

Computing pruned dataset

Each training example is a vertex in the graph. Edges connect pairs of differently-labeled examples $\mathbf{x}$ and $\mathbf{x}'$ whenever $\|\mathbf{x} - \mathbf{x}'\| \leq 2r$

Adversarial Pruning

Bipartite maximum matching via Hopcroft Karp algorithm (1973)

NP-hard when multi-class

Robustness Evaluation

Empirical robustness (ER)

average distance of the correctly predicted $\mathbf{x}$ to the closest $\mathbf{x}_{adv}$

defscore: the ratio of ER w/ and w/o defense

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

- *defscore*: higher the better
- $defscore > 1 \rightarrow$ more robust after defense
- $defscore < 1 \rightarrow$ less robust after defense

average *defscore* over test examples that are correctly predicted

Defense with AP (cont.)

AP improves robustness for non-parametric classifiers

	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**

AT: Madry et al. [11]; Wang's: Wang et al. [15]; RS: Chen et al. [4]

Defense with AP (cont.)

AP improves robustness, but not as good as AT (for parametric classifiers)

	LR		MLP	
	AT	AP	AT	AP
australian	**5.38**	1.24	**8.61**	2.35
cancer	**2.17**	0.99	**2.72**	1.08
covtype	**8.44**	1.56	**11.50**	2.94
diabetes	**4.43**	2.46	**5.31**	2.13
f-mnist06	**3.53**	1.50	**2.59**	2.46
f-mnist35	**3.05**	1.21	**3.94**	1.27
fourclass	**1.80**	1.33	**3.09**	2.35
halfmoon	**1.06**	**1.06**	1.44	**1.47**
mnist17	**3.42**	1.35	**1.51**	1.29

AT: Madry et al. [11]

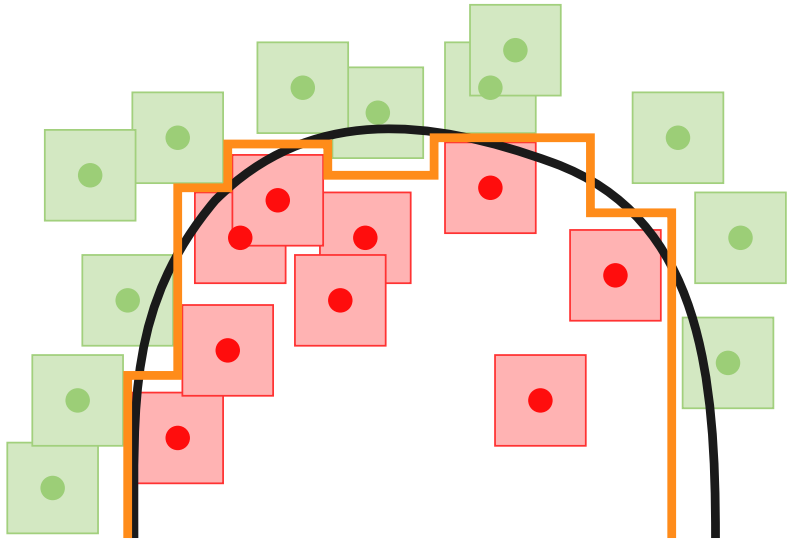
No intrinsic trade-off for r-separated data (cont.)

Lemma

Let $f : \mathcal{X} \to \mathbb{R}^C$ *be a function and consider* $\mathbf{x} \in \mathcal{X}$ *with label* $y \in [C]$. *If*

- f *is* $\frac{1}{r}$*-Locally Lipschitz in a radius* r *around* $\mathbf{x}$, *and*
- $f(\mathbf{x})_j - f(\mathbf{x})_y \geq 2$ *for all* $j \neq y$,

then $g(\mathbf{x}) = \text{argmin}_i f(\mathbf{x})_i$ *is robust at* $\mathbf{x}$ *with radius* r.



No intrinsic trade-off for r-separated data (cont.)

Lemma

Let $f : \mathcal{X} \to \mathbb{R}^C$ be a function and consider $\mathbf{x} \in \mathcal{X}$ with label $y \in [C]$. If

1. *f is $\frac{1}{r}$-Locally Lipschitz in a radius r around $\mathbf{x}$, and*
2. *$f(\mathbf{x})_j - f(\mathbf{x})_y \geq 2$ for all $j \neq y$,*

then $g(\mathbf{x}) = \operatorname{argmin}_i f(\mathbf{x})_i$ is robust at $\mathbf{x}$ with radius r.

Proof.

For any $\mathbf{x}'$ s.t. $\|\mathbf{x}' - \mathbf{x}\|_p \leq 2r$

$$
\begin{aligned}
f(\mathbf{x}')_j &\geq f(\mathbf{x})_j - 1 && f \text{ is } \frac{1}{r}\text{-Locally Lipschitz} \\
&\geq f(\mathbf{x})_y - 1 && (2) \\
&\geq f(\mathbf{x}')_y && f \text{ is } \frac{1}{r}\text{-Locally Lipschitz}
\end{aligned}
$$

Finally, we have $\operatorname{argmin}_i f(\mathbf{x}')_i = \operatorname{argmin}_i f(\mathbf{x})_i = y$ □

No intrinsic trade-off for r-separated data (cont.)

Theorem

If data is r-separated, there always exists a classifier that is perfectly robust and accurate

Key idea

Consider $g(\mathbf{x}) = \text{argmin}_{i \in [C]} f(\mathbf{x})_i$, where $f(\mathbf{x})_i = \frac{1}{r}\text{dist}(\mathbf{x}, \mathcal{X}^{(i)})$.

Show that for every $\mathbf{x} \in \mathcal{X}^{(y)}$ for some $y \in [C]$,

- f is $\frac{1}{r}$-locally Lipschitz in a radius r around $\mathbf{x}$
- $f(\mathbf{x})_j - f(\mathbf{x})_y \leq 2$ for all $j \neq y$

Then base on our lemma, g is robust and accurate.

No intrinsic trade-off for r-separated data (cont.)

Theorem

If data is r-separated, there always exists a classifier that is perfectly robust and accurate

Proof (f is $\frac{1}{r}$-locally Lipschitz).

$f(\mathbf{x})_i = \frac{1}{r}\text{dist}(\mathbf{x}, \mathcal{X}^{(i)})$

$$f(\mathbf{x})_i - f(\mathbf{x}')_i = \frac{\text{dist}(\mathbf{x}, \mathcal{X}^{(i)}) - \text{dist}(\mathbf{x}', \mathcal{X}^{(i)})}{r}$$

$$\leq \frac{\text{dist}(\mathbf{x}, \mathbf{x}')}{r}$$

triangle inequality

□

No intrinsic trade-off for r-separated data (cont.)

Theorem

If data is r-separated, there always exists a classifier that is perfectly robust and accurate

Proof ($f(\mathbf{x})_j - f(\mathbf{x})_y \leq 2$ for all $j \neq y$).

$f(\mathbf{x})_i = \frac{1}{r}\text{dist}(\mathbf{x}, \mathcal{X}^{(i)})$

$$
\begin{aligned}
f(\mathbf{x})_j - f(\mathbf{x})_y &= \frac{\text{dist}(\mathbf{x}, \mathcal{X}^{(j)}) - \text{dist}(\mathbf{x}, \mathcal{X}^{(y)})}{r} \\
&= \frac{\text{dist}(\mathbf{x}, \mathcal{X}^{(j)})}{r} && \text{dist}(\mathbf{x}, \mathcal{X}^{(y)}) = 0 \\
&\geq \frac{\text{dist}(\mathcal{X}^{(y)}, \mathcal{X}^{(j)})}{r} && r\text{-separation} \\
&\geq 2
\end{aligned}
$$

□