# Robustness for Non-Parametric Classification: A Generic Attack and Defense

Yao-Yuan Yang*, Cyrus Rashtchian*, Yizhen Wang and Kamalika Chaudhuri

University of California, San Diego

July 2, 2020

# Introduction (cont.)

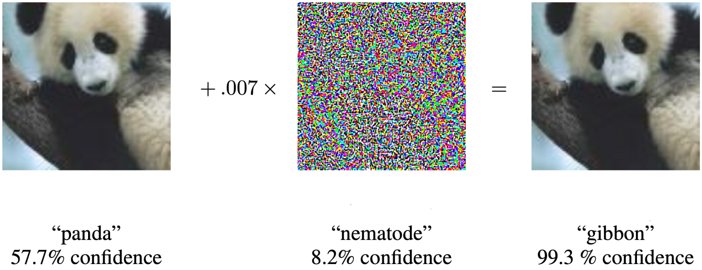



Figure: Goodfellow et al. [4]

**Attack $\mathbf{x}_{adv} = A(f, \mathbf{x}, r)$**

- target classifier $f$
- target example $\mathbf{x}$
- attack budget $r$

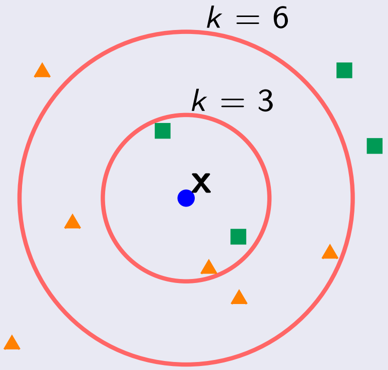
**Optimal attack**

$$\underset{\mathbf{x}_{adv}: f(\mathbf{x}) \neq f(\mathbf{x}_{adv})}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}_{adv}\|_p$$

# Non-parametric Methods

## $k$ nearest neighbor ($k$-NN)

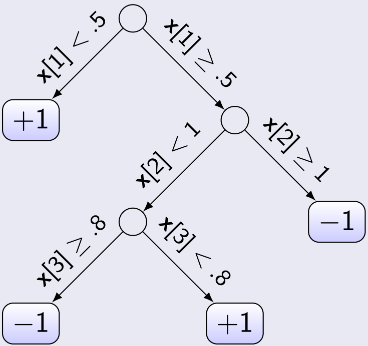
take $k$ closest training examples and output the majority label




## Decision tree and tree ensembles

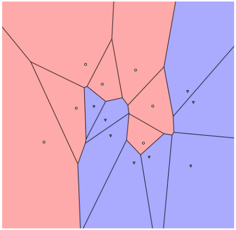
recursively split the data

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

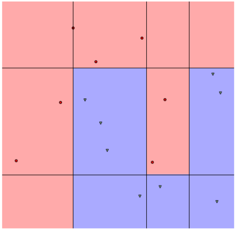


# Region-Based Attack

key observation: decomposition into piece-wise convex regions



(a) 1-NN regions

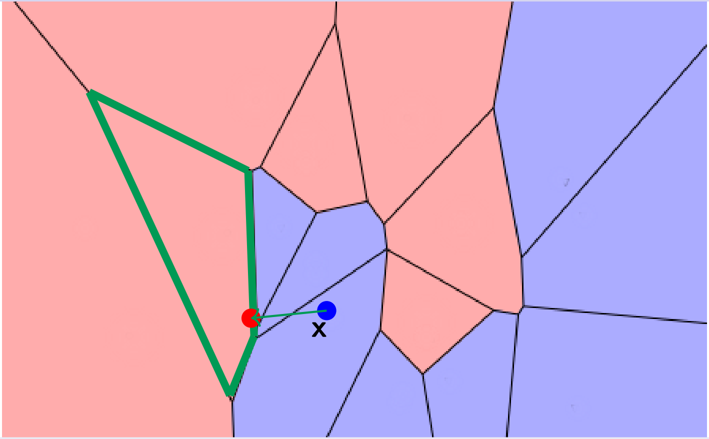


(b) DT regions

**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.

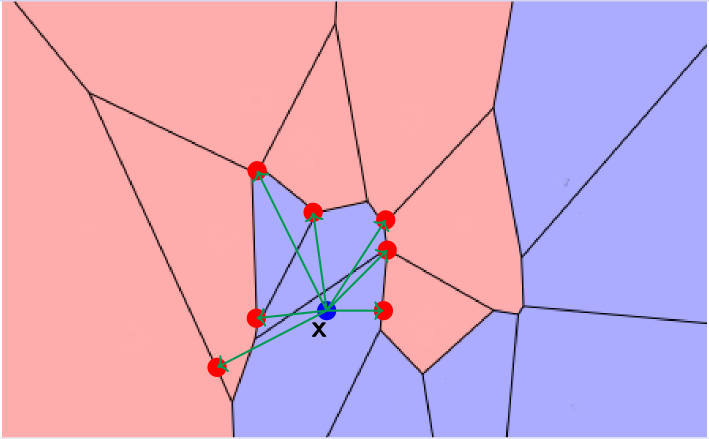
# Region-Based Attack (cont.)




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

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

# Region-Based Attack (cont.)



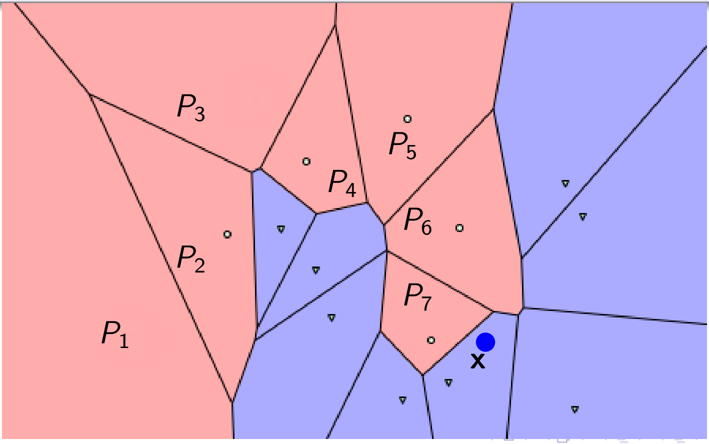

$$\min_{i: f(\mathbf{x}) \neq y_i} \min_{\mathbf{x}_{adv} \in P_i} \|\mathbf{x} - \mathbf{x}_{adv}\|_p$$

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

# Region-Based Attack (Speeding up)

RBA-Approx: consider only a fix number of regions (let's say 3)

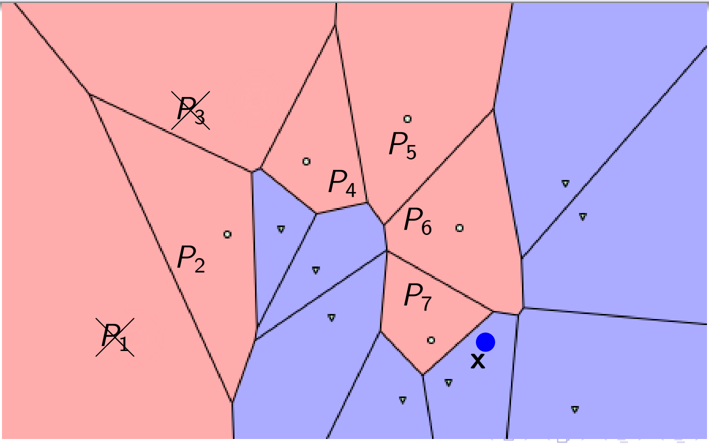
- $P_i$ has at least one training example ($\mathbf{x}_i$) in it (ignore $P_1$, $P_3$)
- sort each region with $\|\mathbf{x}_i - \mathbf{x}\|_p$ (order: $P_7$, $P_6$, $P_5$, $P_4$, $P_2$)
- search only ($P_7$, $P_6$, $P_5$)



# Region-Based Attack (Speeding up)

RBA-Approx: consider only a fix number of regions (let's say 3)

- $P_i$ has at least one training example ($\mathbf{x}_i$) in it (ignore $P_1$, $P_3$)
- sort each region with $\|\mathbf{x}_i - \mathbf{x}\|_p$ (order: $P_7$, $P_6$, $P_5$, $P_4$, $P_2$)
- search only ($P_7$, $P_6$, $P_5$)



# Attack Evaluation

**Empirical robustness (ER)**

average distance of the target example to the adversarial example

average *ER* over test examples that are correctly predicted

*ER*: smaller the better

# Attack Results

| | 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** |

Direct, Kernel: Papernot et al. [8]; BBox: Cheng et al. [3]

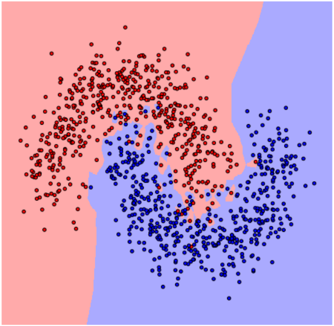
# Attack Results

| | 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 |

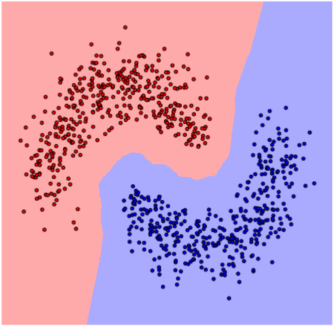
Parernot's: Papernot et al. [8]

Note that Kantchelian et al. [6] also achieves optimal attack on tree-based classifiers

# Defense (motivation)



(c) 1-NN



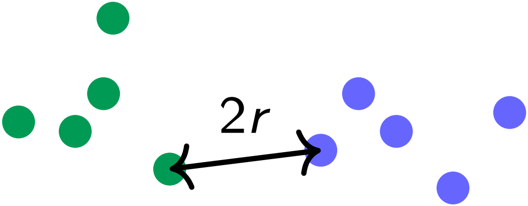
(d) 1-NN with separation (less overlap)

# Adversarial Pruning

a classifier with robust radius $r$ (robust to attacks with an attack budget $r$)

**Defense strategy**

1. remove minimum # of examples s.t. distance between differently-labeled examples are $\geq 2r$ (minimum vertex cover problem)
2. learn a non-parametric classifier on the modified dataset



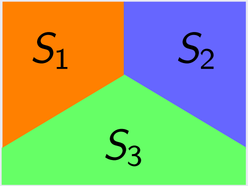

Next, some theoretical justifications

Similar technique has been used by Gottlieb et al. [5] for the consistency of 1-NN, but not for robustness

# Adversarial Pruning ($r$-Optimal Classifier)

## Bayes-optimal classifier

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

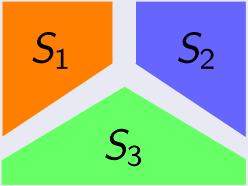



## $r$-Optimal classifier

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

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

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



# Defense Evaluation

## Recall: Empirical robustness

average distance of the target example to the adversarial example

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

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

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

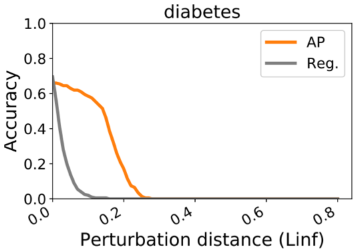
average *defscore* over test examples that are correctly predicted

# Defense Results

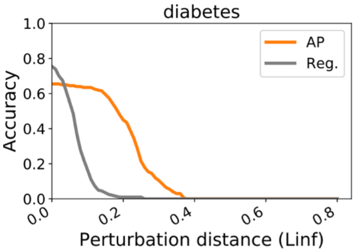
| | 1-NN | | | 3-NN | | DT | | | RF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AT | Wang's | AP | AT | AP | AT | RS | AP | AT | RS | AP |
| australian | 0.64 | **1.65** | **1.65** | 0.68 | **1.20** | 2.36 | **5.86** | 2.37 | 1.07 | **1.12** | 1.04 |
| cancer | 0.82 | 1.05 | **1.41** | 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. [7]; Wang's: Wang et al. [9]; RS: Chen et al. [2]

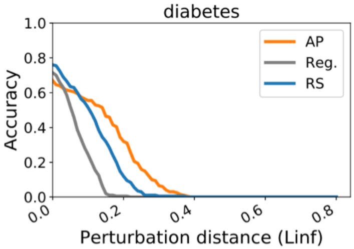
# Defense Results (cont.)



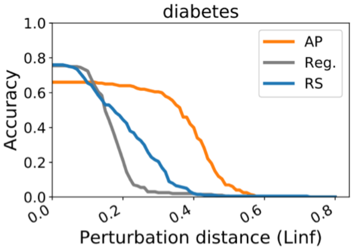

(e) 1-NN




(f) 3-NN




(g) Decision tree




(h) Random forest

# Conclusion

- an attack algorithm based on decomposing feature space into convex regions then attack each region independently
- a defense algorithm by modifying the dataset so the dataset is more separated
- $r$-Optimal classifier as a robust analog to the Bayes optimal classifier

**future work**

- some more classifier specific attack/defense algorithm
- $r$-Optimal classifier (Bhattacharjee and Chaudhuri [1])

Thank you for listening.

## More information

- Paper: `https://arxiv.org/abs/1906.03310`
- Code: `https://git.io/JfyXo`
- Blog: `https://ucsdml.github.io/`

## Contact

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

# References I

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

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

[3] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient Hard-label Black-box Attack: An Optimization-based Approach. In *ICLR*, 2019.

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

[5] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

# References II


[6] Alex Kantchelian, JD Tygar, and Anthony Joseph. Evasion and Hardening of Tree Ensemble Classifiers. In *ICML*, pages 2387–2396, 2016.

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

[8] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

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