Fairness Through Robustness:
Investigating Robustness Disparity in Deep Learning

Vedant Nanda*, **Samuel Dooley***, Sahil Singla, Soheil Feizi, John P. Dickerson

9 March 2021

FAccT 2021

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

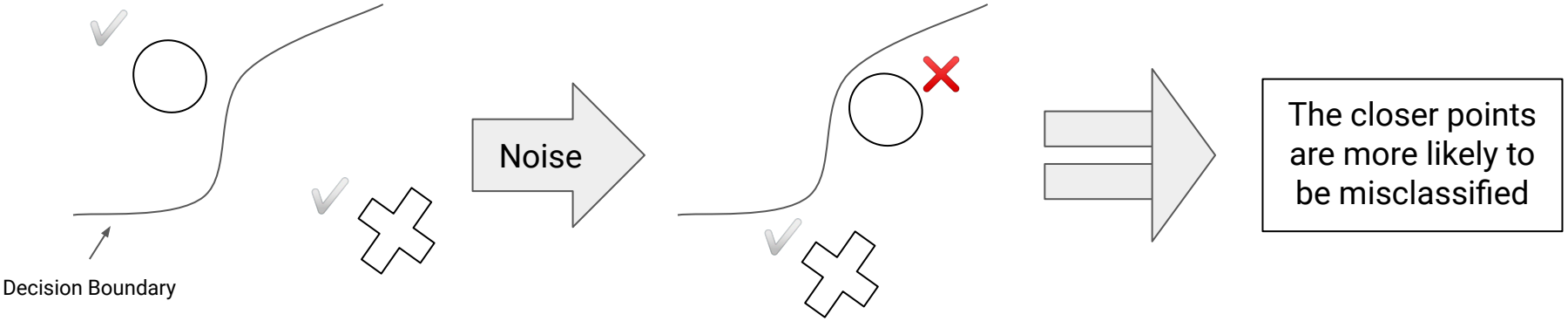
ROBUSTNESS BIAS

Most notions of fairness are defined by system outputs

Noise can cause a system to treat groups unfairly

Sneak Peak: YES!

Are some types of points <u>systematically more robust</u> than others?



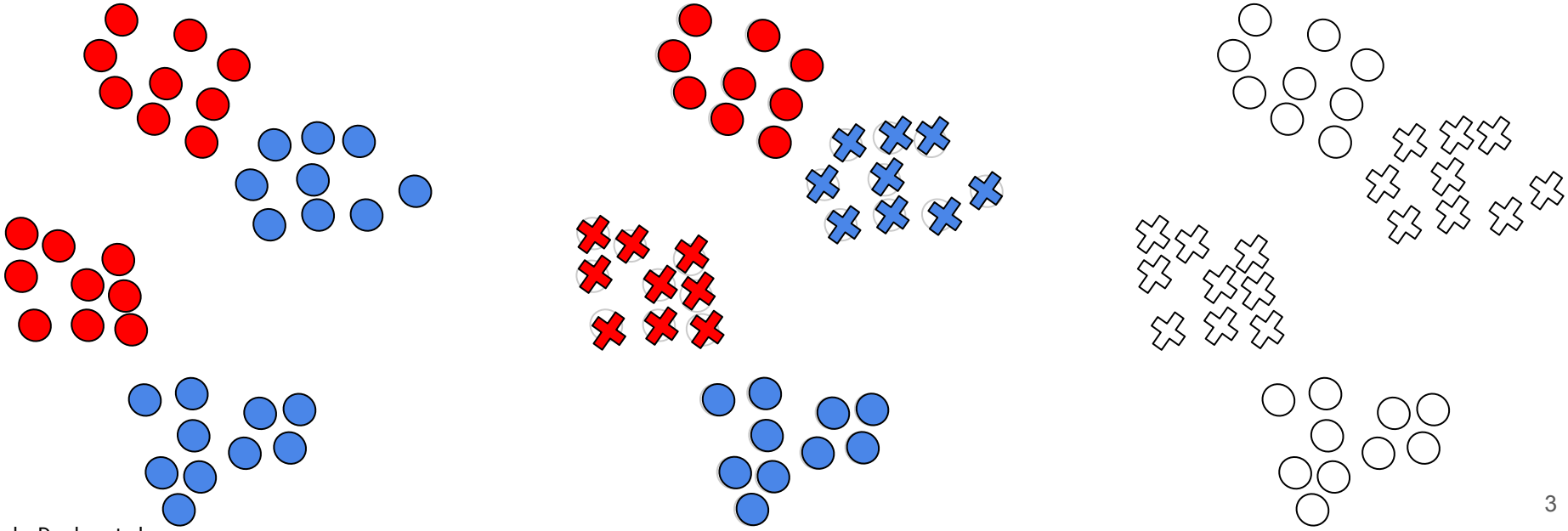
Nanda, Dooley et al.

DATA PRELIMINARIES – PARTITION

Partition by
Predictor Variable

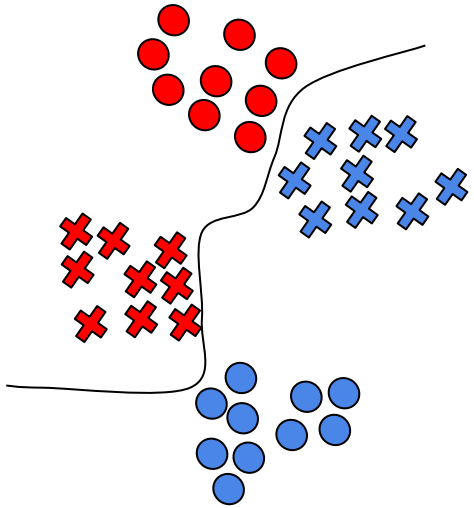
Dataset

Partition by
Sensitive Attribute



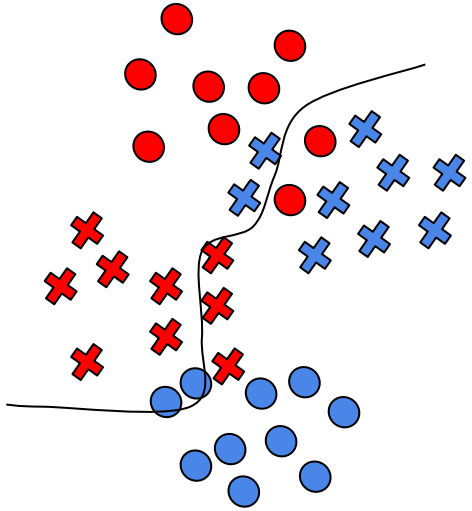
Nanda, Dooley et al.

INTUITION



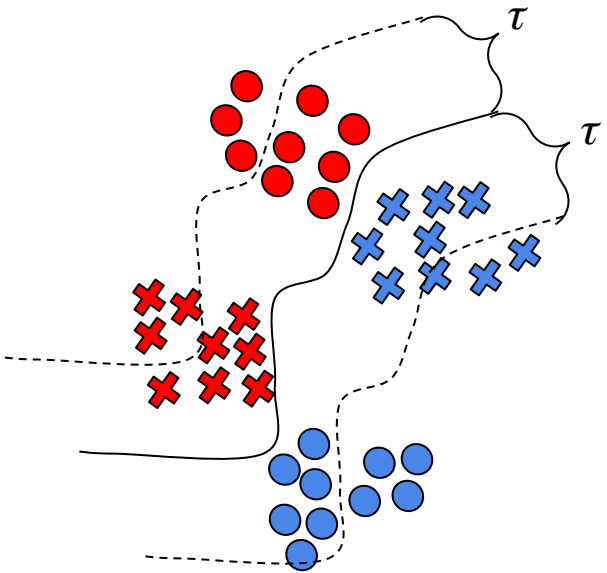
Nanda, Dooley et al.

INTUITION



Nanda, Dooley et al.

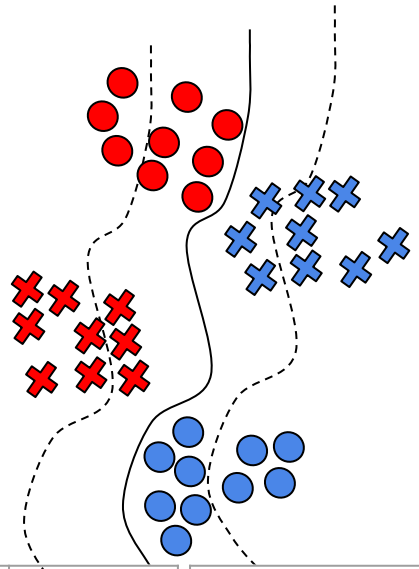
INTUITION



Predictor Partition	Robustness	Sensitive Partition	Robustness
Red	2/3	Circles	2/3
Blue	2/3	X's	2/3

ROBUSTNESS BIAS

Depends on the
Data, Model, and Partition
***and also on tau**



Predictor Partition	Robustness	Sensitive Partition	Robustness
Red	1/2	Circles	2/3
Blue	1/2	X's	1/3

Nanda, Dooley et al.

METRICS

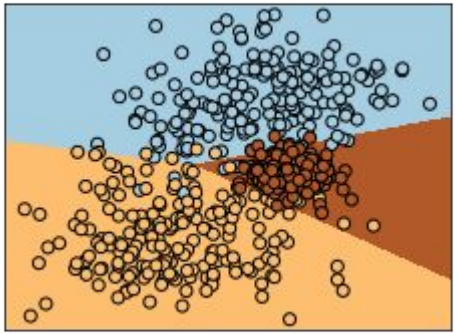
- Given a dataset D, $\tau > 0$, and partition $P \subset D$.
- Let $d_\theta(x)$ be the minimal distance from point x to a decision boundary.
- Calculate

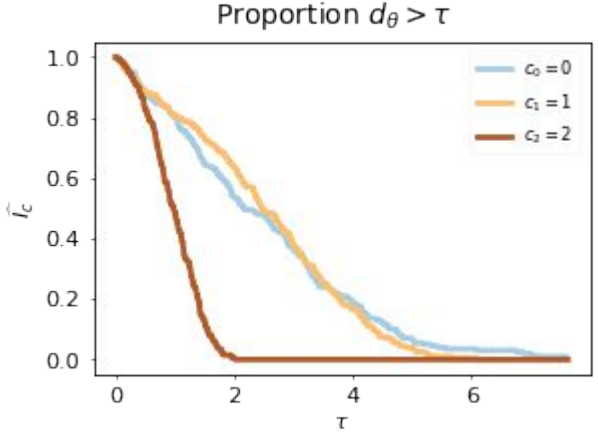
Proportion of that are more than tau from decision boundary

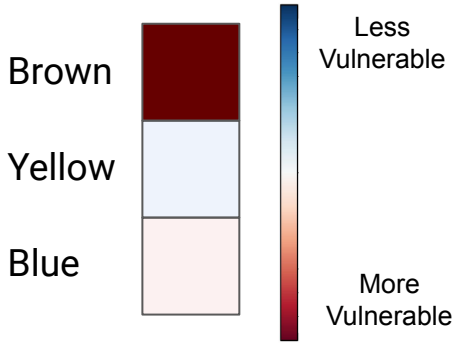
$$\widehat{I_P}(\tau) = \frac{|\{(x,y) \in P \mid d_\theta(x) > \tau, y = \hat{y}\}|}{|P|}$$

Relative average distance to the decision boundary

$$\sigma(P) = \frac{AUC(\widehat{I_P}) - AUC(\sum_{P' \neq P} \widehat{I_{P'}})}{AUC(\sum_{P' \neq P} \widehat{I_{P'}})}$$







Nanda, Dooley et al.

RESEARCH QUESTIONS

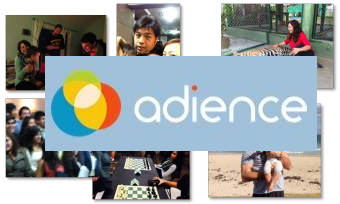
- Does robustness bias exist in the wild?
- How much does it depend on the dataset? the *partition*? the **model**?

Class

Class

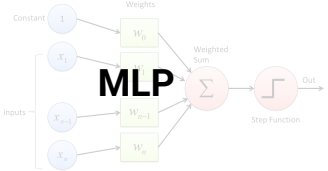
Age

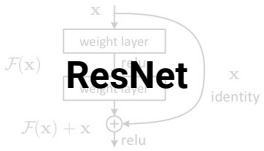
Gender, Race

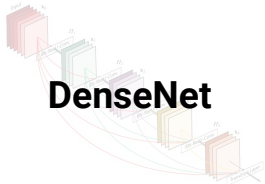


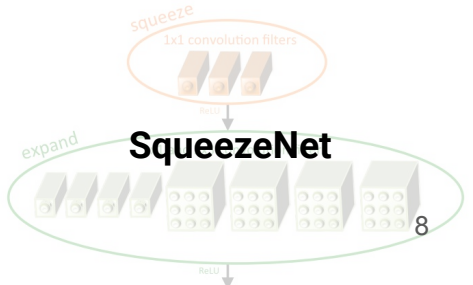
Age

Gender





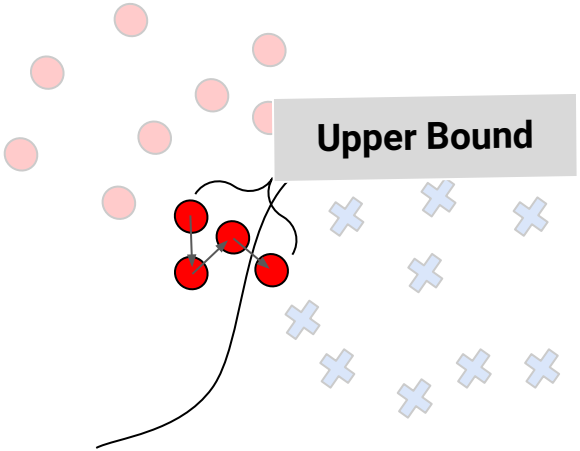




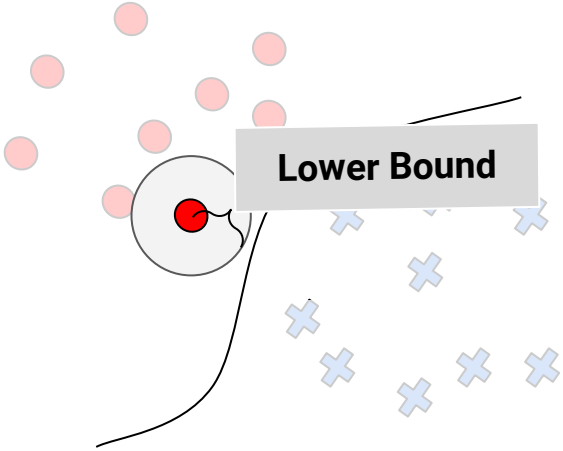
Nanda, Dooley et al.

BUT WE CAN'T COMPUTE $d_\theta(x)$

- All our models (except MLP) do not permit direct computation
- We can approximate $d_\theta(x)$ with upper or lower bounds



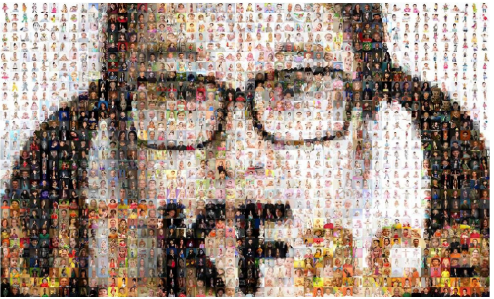
DeepFool and CarliniWagner



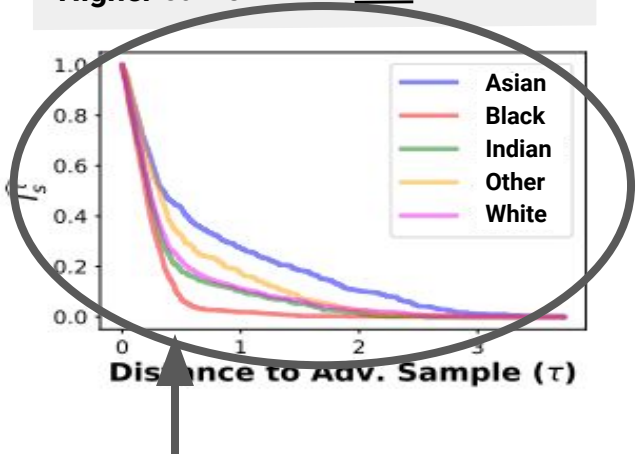
Randomized Smoothing

Nanda, Dooley et al.

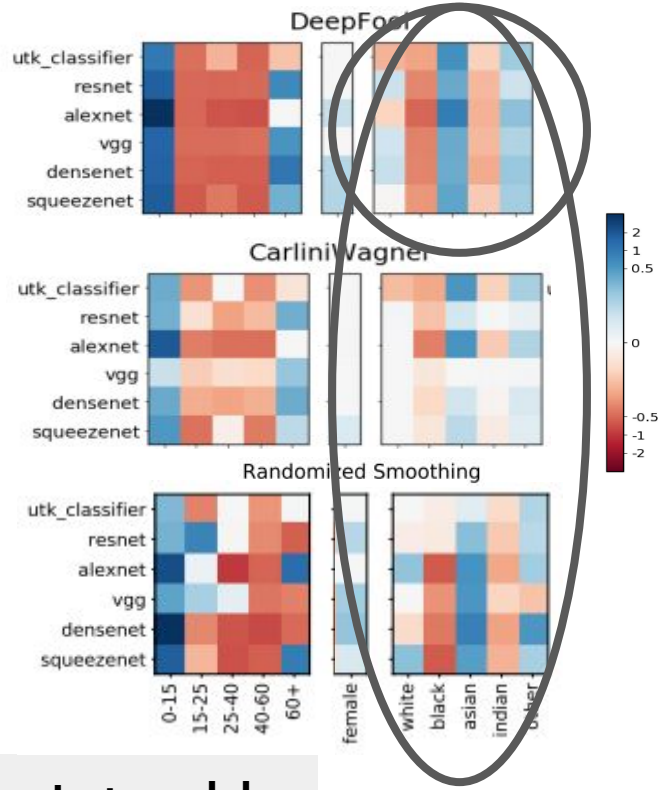
RESULTS



RECALL:
Lower curve means <u>more</u> vulnerable
Higher curve means <u>less</u> vulnerable



RECALL:
Red means partition is <u>more</u> vulnerable
Blue means partition is <u>less</u> vulnerable

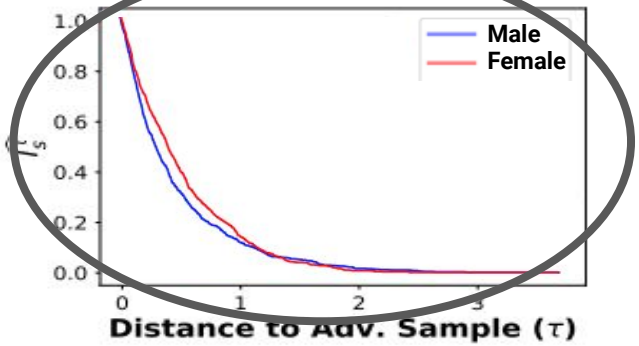


Qualitatively similar behavior for other datasets + models

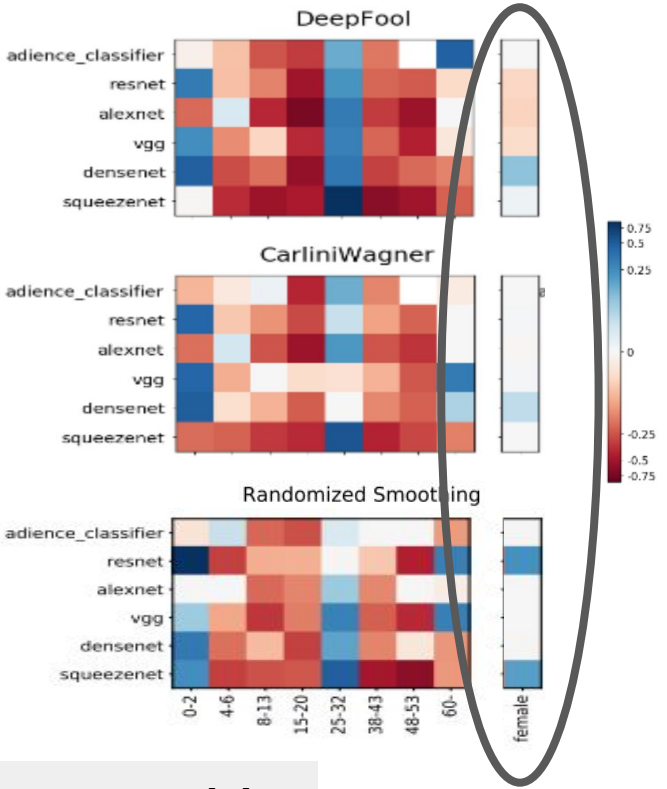
Nanda, Dooley et al.

RESULTS

RECALL:
Lower curve means more vulnerable
Higher curve means less vulnerable



RECALL:
Red means partition is more vulnerable
Blue means partition is less vulnerable



Qualitatively similar behavior for other datasets + models

Nanda, Dooley et al.

RESULTS

Our results suggest that:

- Robustness Bias appears to be present in the wild.
- Some classes appear to uniformly be more/less robust across all models.
- Some partitions of datasets appear to exhibit robustness bias more than others.
- DeepFool and CarliniWagner agree on the sign of the bias (with $p<0.001$ in the Pearson's Chi-squared).
- There is some evidence of agreement between the upper and lower bounds.

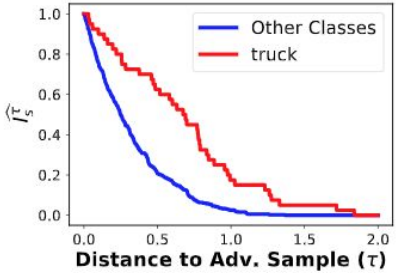
CAN WE MITIGATE ROBUSTNESS BIAS?

Attempted the obvious regularization fix.

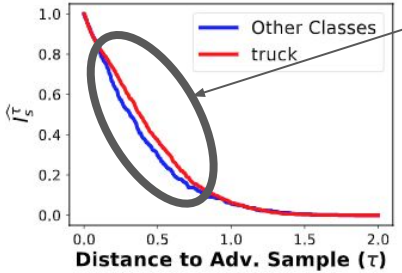
$$\left| \frac{1}{\sum_{x \notin P} \mathbb{1}\{y = \hat{y}\}} \sum_{\substack{x \notin P \\ y = \hat{y}}} \mathbb{1}\{d_\theta(x) > \tau\} - \frac{1}{\sum_{x \in P} \mathbb{1}\{y = \hat{y}\}} \sum_{\substack{x \in P \\ y = \hat{y}}} \mathbb{1}\{d_\theta(x) > \tau\} \right|$$

Tradeoff in equal treatment vs efficiency in general

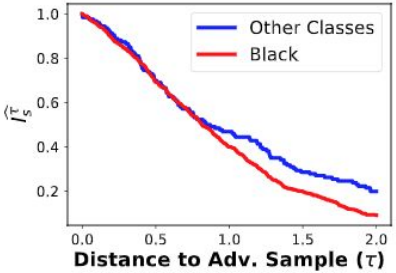
More equal, but less robust



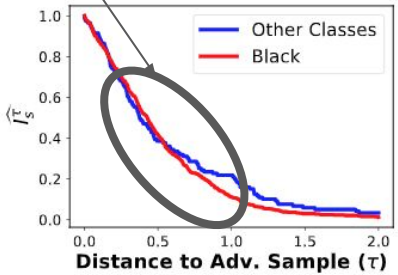
Without Regularization



With Regularization



Without Regularization



With Regularization

Nanda, Dooley et al.

FUTURE DIRECTIONS

- Use adversarial training to balance mitigation with overall robustness hit
- What about on deployed models?
  - Quantifying harms - ethical concerns for such measurements makes it hard
- How do other methods to mitigate fairness affect robustness bias?
- How does robustness bias change for robust models?
- Benchmarking current models for other kinds of perturbation based biases
  - Synthetic corruptions eg: ImageNet-C
  - Natural corruptions eg: ImageNetV2

Check out our paper!

https://arxiv.org/abs/2006.12621

Get in touch: {vedant,sdooley1}@cs.umd.edu

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS