Fair-MAML trains fair meta-models that can be fine-tuned for specific tasks with minimal data.

Fair Meta-Learning: Learning How to Learn Fairly

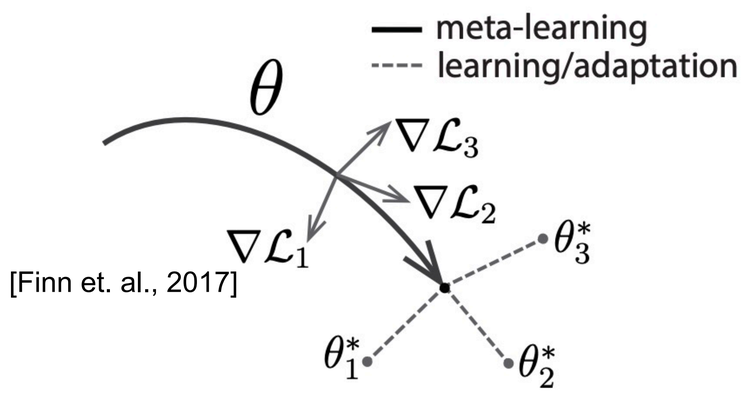
Dylan Slack, Sorelle Friedler, and Emile Givental

MOTIVATION

- Minor changes in test distribution can have significant effects on fairness (see Fairness Warnings). How can we train a model that copes?

METHODS

- We can train *fair meta-model* that contains general features relating to both fairness and accuracy using *model agnostic meta-learning* with added fairness objective (Fair-MAML).
- Fair-MAML can be fine-tuned to new fairness tests to achieve high degrees of accuracy with minimal data.



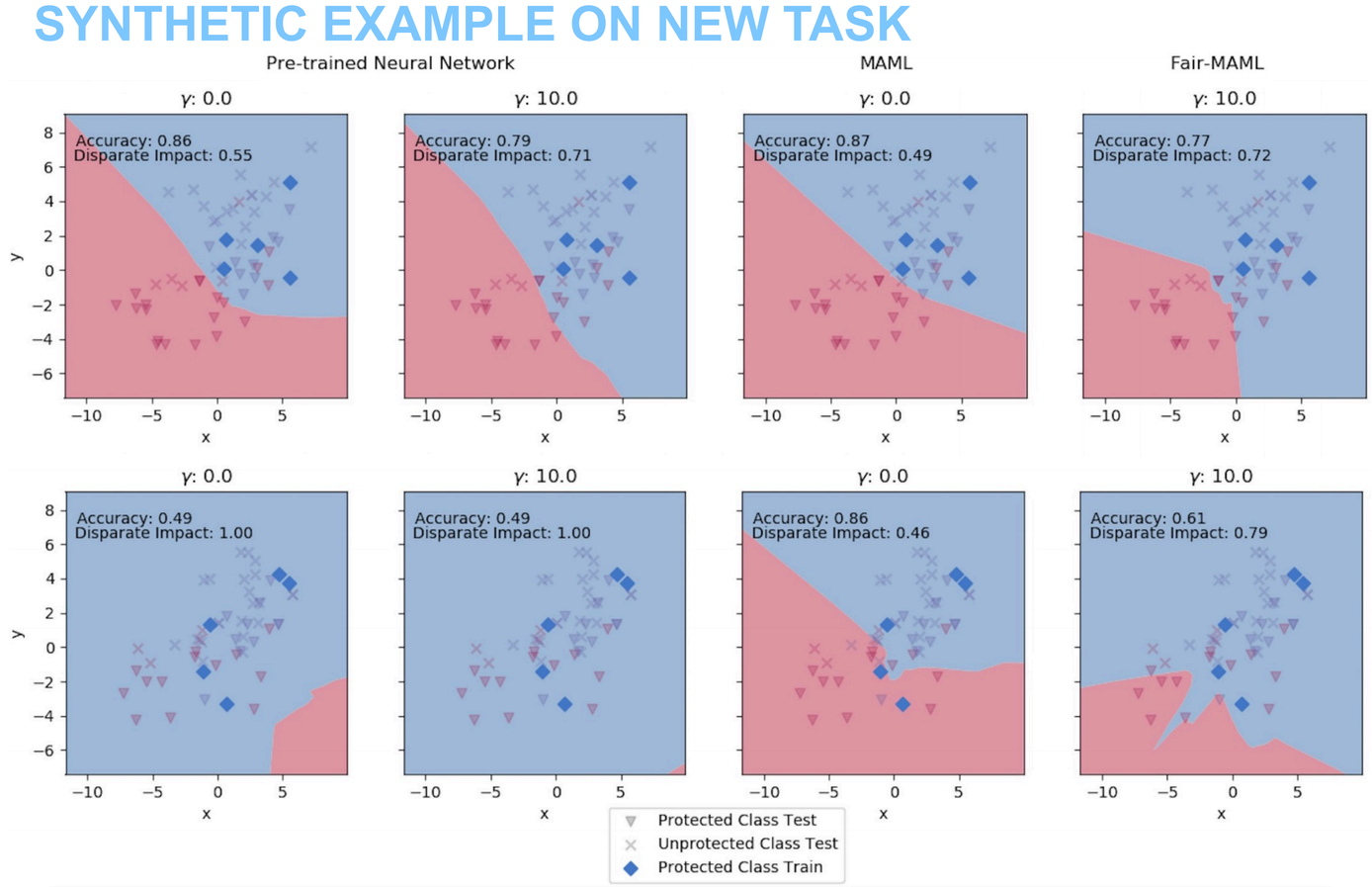
[Finn et. al., 2017]

DEMOGRAPHIC PARITY REGULARIZER IN TASK LOSS:

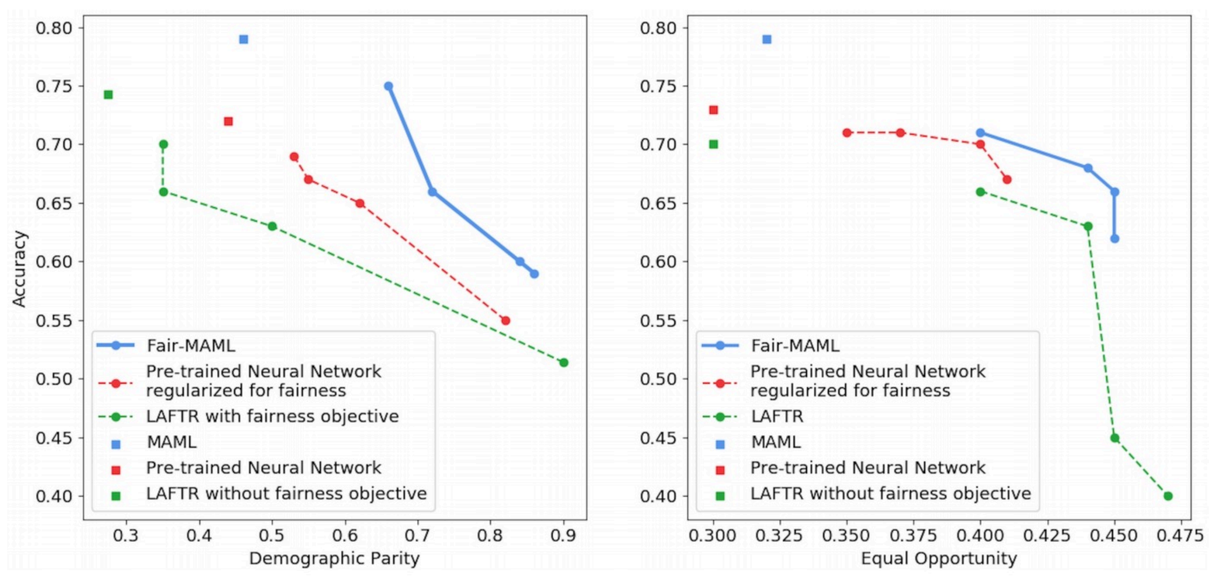
- (D_0 indicates protected instances)

$$\mathcal{R}_{dp}(f_\theta, \mathcal{D}) = 1 - P(\hat{Y} = 1 | A = 0)$$
$$\approx 1 - \frac{1}{|\mathcal{D}_0|} \sum_{x \in \mathcal{D}_0} P(f_\theta(x) = 1)$$

SYNTHETIC EXAMPLE ON NEW TASK



COMMUNITIES AND CRIME EXAMPLE



HAVERFORD COLLEGE

UCI University of California, Irvine

Check us out at FAT* 2020:

Fairness Warnings & Fair-MAML: Learning Fairly from Minimal Data

Scan code for arXiv.