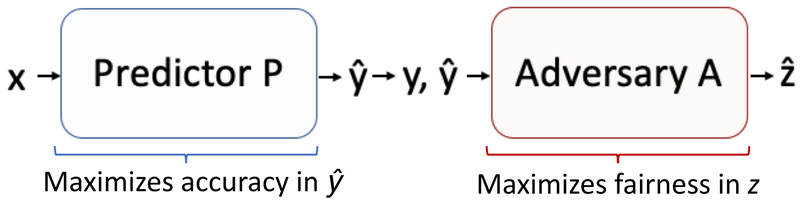
FAD: Fairness through Adversarial Discrimination

Alexandra Henzinger (ahenz), Justin Chen (jyc100)

Motivation

- ML models mirror biases in data: model predicts $\hat{y}$ based on features *x*, containing protected feature z
- Goal: augment models to induce **fair** predictions
- Numerous definitions of *fairness*:
  - **Demographic parity**: $\hat{y}$ and *z* independent
  - **Equality of opportunity**: $\hat{y}$ and *z* conditionally independent given y = 1
  - **Equality of odds**: $\hat{y}$ and *z* conditionally independent given y
- Incongruent: one model can never satisfy all 3 definitions
- **Adversarial network** to encode fairness into model:



- **Post-training processing**: enforce fairness on black-box model by ROC analysis, e.g. equating true positive and false positive rates [Hardt et al. (2016)]

What are the tradeoffs between fairness and accuracy across methods?

Data + Features

- **UCI Adult income dataset**: predicting income (≤/>$50K) based on demographic census data of individuals (*age*, *sex*, *race*, *workclass*, *occupation*, *investments*, *education degree*, *marital status*, *relationship*, *native country*)
- **Protected variables *z*** : sex (presented here), race, age
- 32K individuals in train & 16K in test

PROTECTED VARIABLE DISTRIBUTION IN UCI INCOME DATA

Protected Var	Value	Distribution
sex	Male	67%
	Female	33%

Adversarial Model

$$L_P = L_{CE}(y, \hat{y}) - \alpha L_A(z, \hat{z})$$

Predictor loss — Predictor logistic loss — Adversary logistic loss

- **Hyperparam α** regulates accuracy/fairness tradeoff
- Input to adversary A depends on choice of fairness metric: y for demographic parity; (y, ŷ) for equality of odds/opp.
- 3-layer neural networks for predictor and adversary.

Post-Processing

- Alter class-specific thresholding of logits for predictions to align TP and FP across all classes z [Hardt et al. (2016)]
- **True positive (TP) rate**: $p(\hat{y} = 1 | y = 1, z)$
- **False positive (FP) rate**: $p(\hat{y} = 1 | y = 0, z)$
  - → Equivalent TP across all z gives equality of opportunity
  - → Equivalent TP and FP across all z gives equality of odds
  - → Equivalent $p(\hat{y} = 1 | z)$ across all z gives parity
- **EO** post-processing enforces equality of opportunity; **DP** enforces demographic parity.

Results

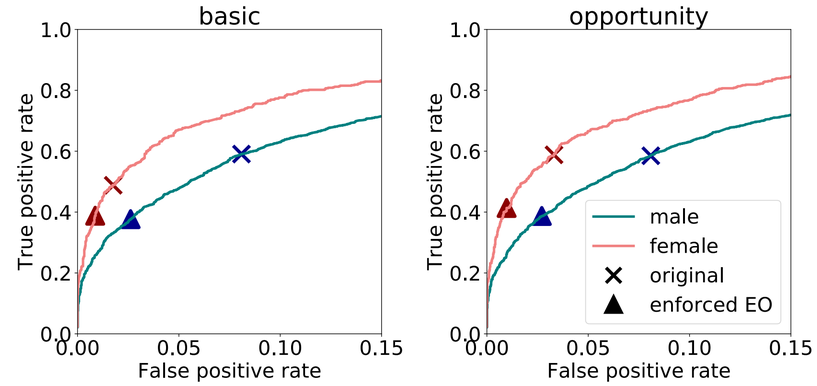
Method	Accuracy	Parity-Gap	FP-Gap	TP-Gap
Basic	**85.7**	0.16	0.06	0.10
Opportunity ($\alpha = 10$)	85.4	0.14	0.47	**0.00**
Odds ($\alpha = 10$)	85.0	0.11	**0.03**	**0.03**
Parity ($\alpha = 10$)	83.9	**0.02**	**0.03**	0.29
Beutel [1]	82.3	0.19	0.15	0.07
Zhang [7]	84.5	–	**0.01**	**0.01**
Basic + *EO*	85.5	0.16	0.06	**0.02**
Opportunity + *EO*	85.5	0.16	0.07	**0.02**
Basic + *DP*	83.6	**0.01**	0.04	0.29
Parity + *DP*	83.6	**0.00**	0.04	0.31

References

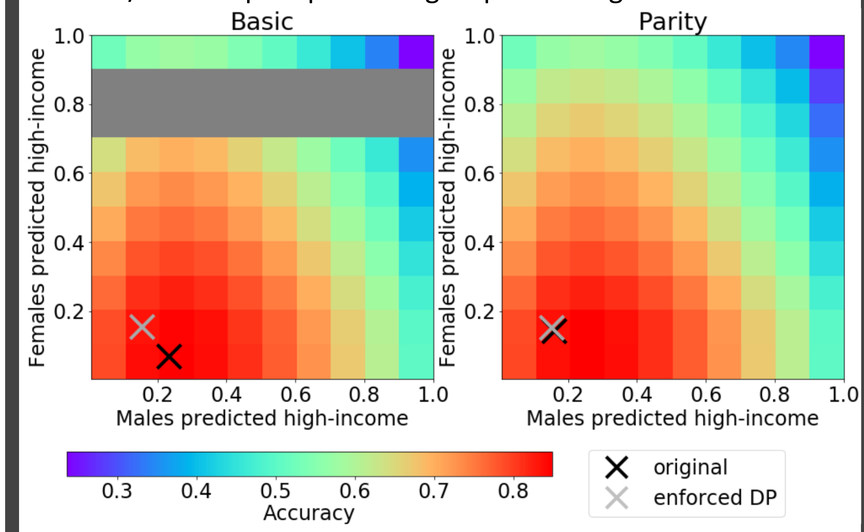
M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. ArXiv e-prints, October 2016.
C. Wadsworth, F. Vera, and C. Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. FAT/ML, 2018.
B. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. CoRR, abs/1801.07593, 2018.

Discussion

- *Fig 1*: ROC Curve comparison for basic and opportunity models, with/without post-processing step enforcing EO



- *Fig 2*: Accuracy comparison for basic and parity models, with/without post-processing step enforcing DP



- Adversarial models show **initial gains in fairness** with little loss in accuracy.
- Post-processing techniques create predictions that are **approximately as fair**, using basic model (no adversary).

Future Work

- Expand to more datasets & protected variables
- Test different predictor and adversary architectures
- Incorporate fairness definition directly into differentiable loss function, to train only a predictor without the adversary (e.g. via Lagrange multipliers)