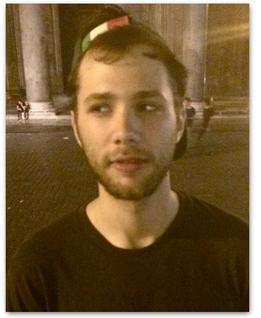
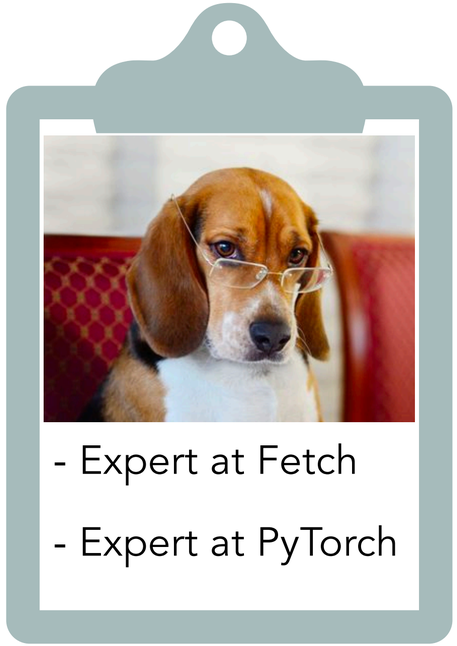
THE IMPLICIT FAIRNESS CRITERION OF UNCONSTRAINED LEARNING

Lydia T. Liu (UC Berkeley)



Joint work with **Max Simchowitz** and **Moritz Hardt**

A TYPICAL APPLICATION OF MACHINE LEARNING TODAY



You're Hired!

$X \in \mathcal{X}$ feature space

Resume

Ads

Movies

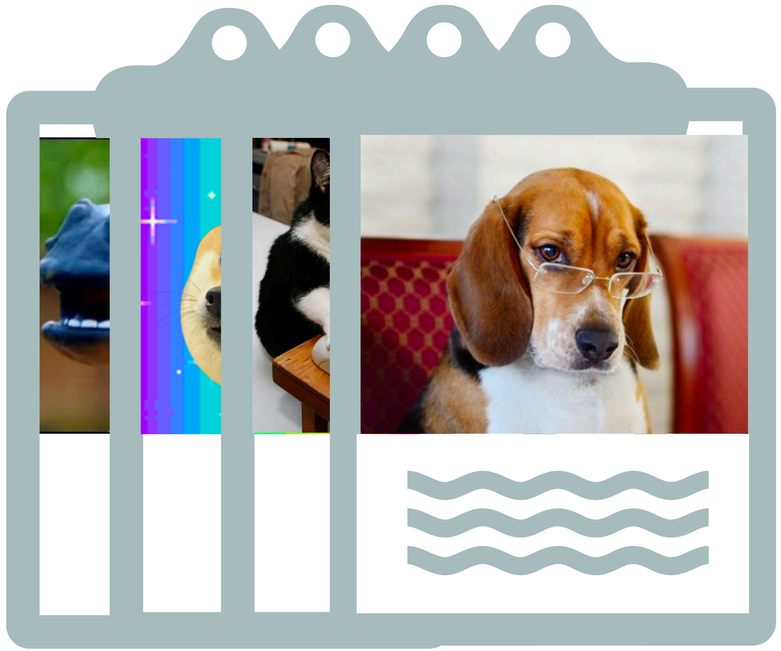
$Y \in \{0, 1\}$ outcome

Hiring Decision

User clicks

User likes

A TYPICAL APPLICATION OF MACHINE LEARNING TODAY



$X_i \in \mathcal{X}$

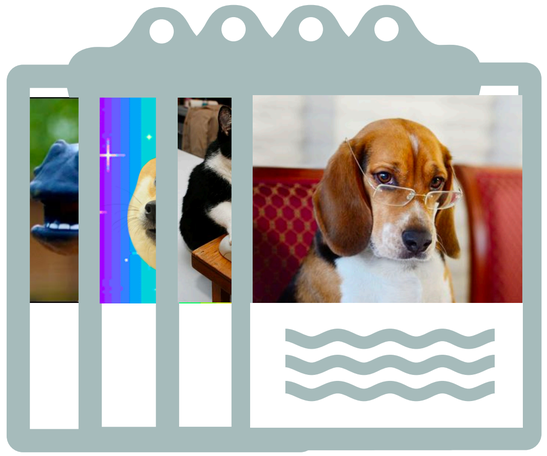
$Y_i \in \{0, 1\}$

Learn a score function $f(X)$ that is "close" to Y

$$f^U(x) = \mathbb{E}[Y \mid X = x]$$

Optimal score—minimizes MSE (also other losses) for each x

A TYPICAL APPLICATION OF MACHINE LEARNING TODAY



$X_i \in \mathcal{X}$

$Y_i \in \{0, 1\}$

$f^U(x) = \mathbb{E}[Y \mid X = x]$ satisfies **Calibration**

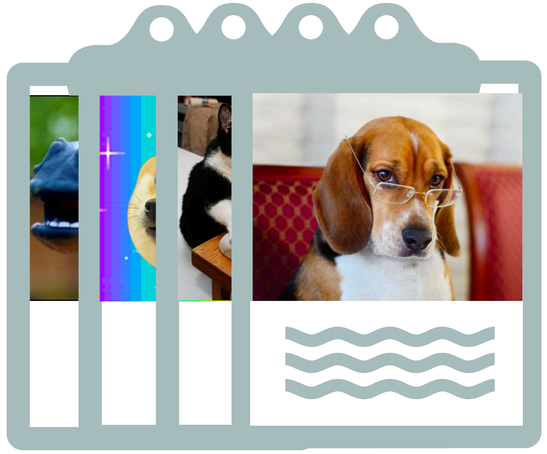
$$\mathbb{E}[Y \mid f(X) = c] = c$$

[Cox, 1958, Murphy and Winkler, 1977, Dawid, 1982, DeGroot and Fienberg, 1983, Platt, 1999, Zadrozny and Elkan, 2001, Niculescu-Mizil and Caruana, 2005]

Score function corresponds to probability that $Y = 1$

A TYPICAL APPLICATION OF MACHINE LEARNING TODAY

Application: proposed as criterion for **fairness** when A is a **group**



$X_i \in \mathcal{X}$

$Y_i \in \{0, 1\}$

Calibration w.r.t. A

$$\mathbb{E}[Y \mid f(X) = c, A = a] = c$$

Calibration

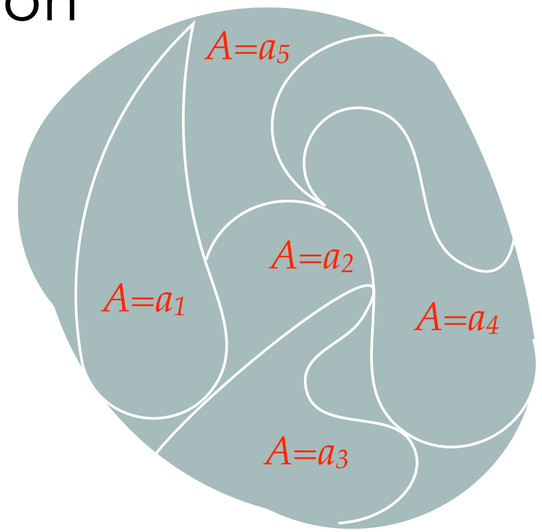
$$\mathbb{E}[Y \mid f(X) = c] = c$$

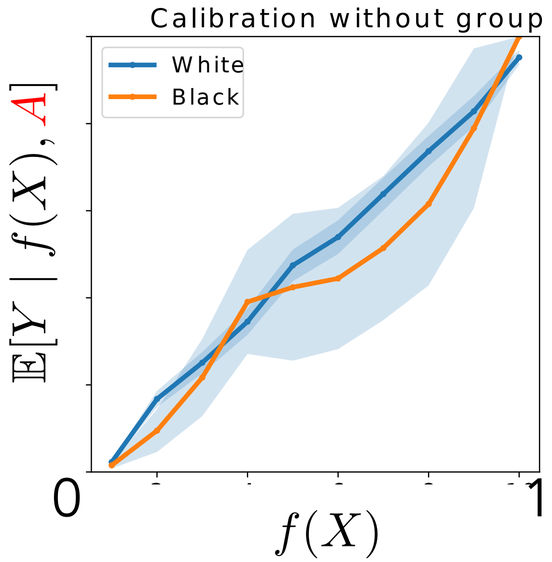
Consider A, an attribute that may or may be in X

GROUP CALIBRATION

$$\mathbb{E}[Y \mid f(X) = c, A = a] = c$$ **Calibration** w.r.t. A

Population



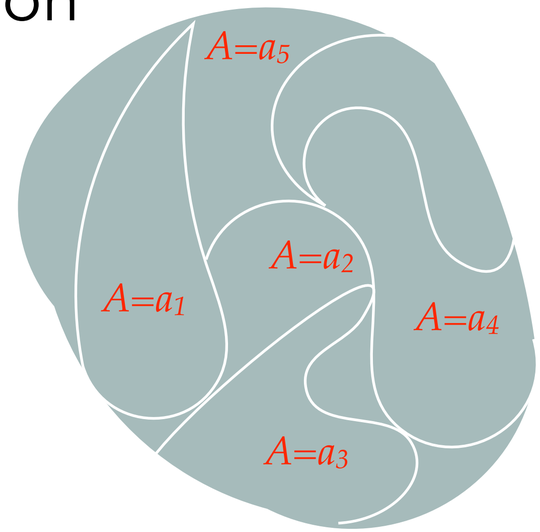


e.g. A is race

GROUP CALIBRATION

$$\mathbb{E}[Y \mid f(X) = c, A = a] = c$$ **Calibration** w.r.t. A

Population



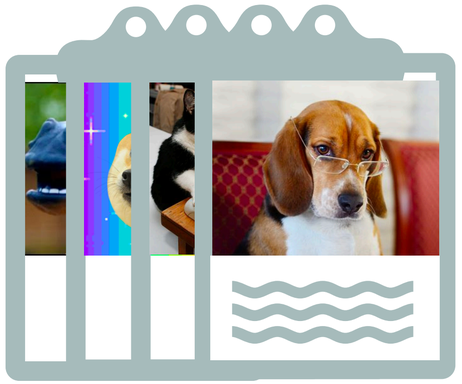
The **Calibrated Bayes Score**

$$f^B(x, a) = \mathbb{E}[Y \mid X = x, A = a]$$

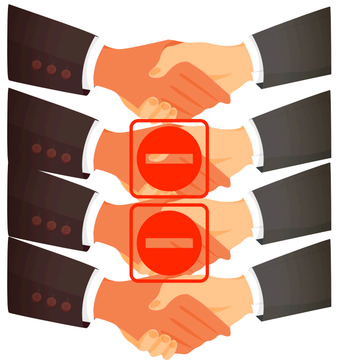
satisfies **Calibration** w.r.t. A

$$\mathbb{E}[Y \mid f^B(X, A), A] = f^B(X, A)$$

TYPICAL APPLICATION OF MACHINE LEARNING TODAY



$X_i \in \mathcal{X}$



$Y_i \in \{0, 1\}$

A form of ***unconstrained machine learning*** (no constraints based on A)

"Status Quo": Empirical Risk Minimization (ERM)

1. Specify a model class $\mathcal{F}$
2. Learn a score function $\hat{f} \in \mathcal{F}$ that minimizes prediction loss over data $(X_i, Y_i)_{i=1}^n$

X may not contain A

Q: When is ERM **calibrated** w.r.t. A?

THIS WORK

Unconstrained machine learning via ERM is a simple recipe for achieving group calibration w.r.t. A, provided that

1. the function class $\mathcal{F}$ is sufficiently rich,
2. there are enough training samples, and
3. the available features $\boldsymbol{X}$ can approximately capture the group attribute A for purposes of predicting $\boldsymbol{Y}$

RESULT: UPPER BOUND ON CALIBRATION GAP

- **Calibration gap** of score f

$$\mathbf{cal}_f(A) := \mathbb{E}\left[\left|f(X) - \mathbb{E}[Y \mid f(X), A]\right|\right]$$

- Consider a loss function $\ell = \ell(f(X), Y)$ (e.g. square or logistic loss).
  The ***risk*** of the score is the average loss over the population distribution:

$$\mathcal{L}(f) = \mathbb{E}[\ell(f(X), Y)]$$

- Our main result relates the calibration gap of a score to its *excess* risk compared to the **Calibrated Bayes Score** $f^B(x, a) = \mathbb{E}[Y \mid X = x, A = a]$

$$\mathcal{L}^* = \mathbb{E}[\ell(f^B(X, A), Y)]$$

Calibrated Bayes Risk

RESULT: UPPER BOUND ON CALIBRATION GAP

- **Theorem 1.** For a broad class of loss functions that includes the square loss and logistic loss, we have

$$\mathbf{cal}_f(A) \leq \mathcal{O}\left(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}\right)$$

  - Big O for constants that depend only on the loss function ℓ
  - Corresponding **lower bound** shows that the square-root relationship between excess risk and **calibration** gap is tight in the worst case.
- Any score with **small excess risk** over the calibrated Bayes risk will be well-**calibrated** with respect to the group attribute A

IMPLICATIONS OF THEOREM 1

- Given a dataset of size n sampled from the population distribution, a natural strategy for achieving group **calibration** is
  - the unconstrained **empirical risk minimization (ERM)** over a model class $\mathcal{F}$

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \ .$$

- The risk of $\hat{f}_n$ converges in probability to the least risk of any score function in the class, $\min_{f \in \mathcal{F}} \mathcal{L}(f)$.

IMPLICATIONS OF THEOREM 1

- It is possible for ERM to attain small excess risk relative to the calibrated Bayes risk even if the group attribute A is **not** in the training dataset.
- **Example.** Let $\ell(z, y) := (z - y)^2$ denote the square loss. Then we can decompose the excess risk as follows

$$\mathcal{L}(\hat{f}) - \mathcal{L}^* = \left(\mathcal{L}(\hat{f}) - \min_{f \in \mathcal{F}} \mathcal{L}(f)\right) + \left(\min_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}(f^U)\right) + \mathbb{E}_X[\mathrm{Var}_A[f^B \mid X]$$

vanishes at $1/\sqrt{n}$ rate by ERM

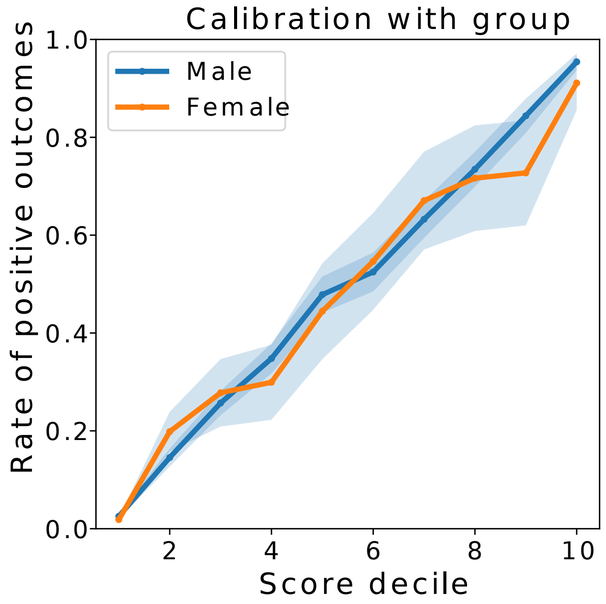
flexibility of the function class

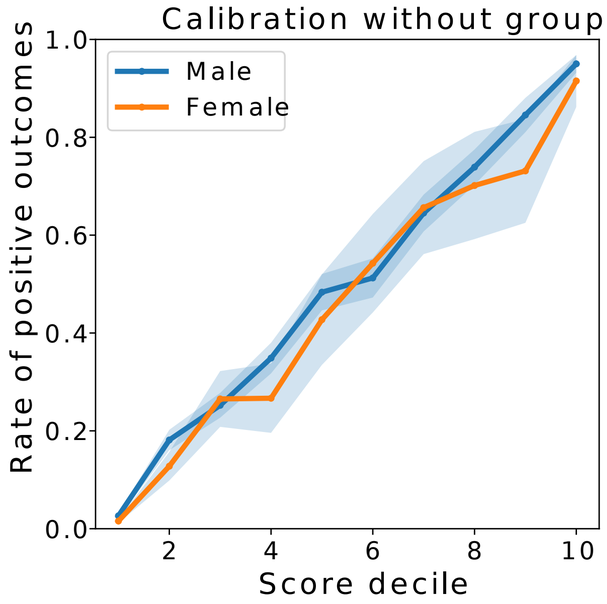
$$f^U = \mathbb{E}[Y \mid X]$$

Conditional variance of f^B given X; 0 if X captures all necessary information about A

EXPERIMENTS ON UCI ADULT

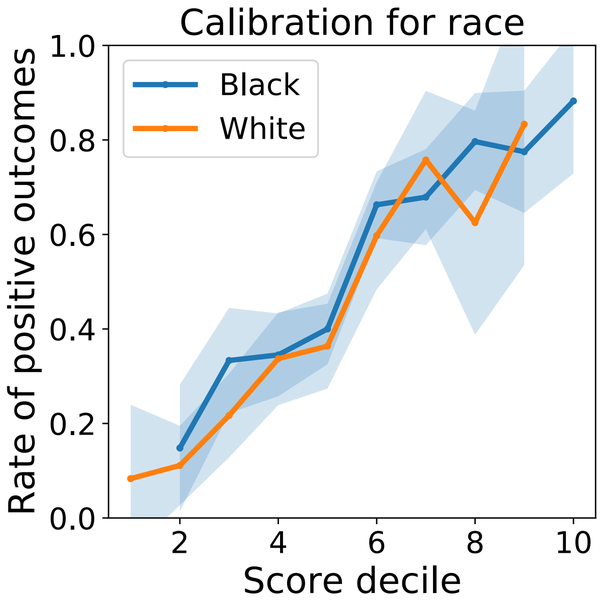
- 14 features, 48842 individuals, predict if annual income > $50,000
- **Observation 1**: ERM score is close to **calibrated** by group even when trained without the group attribute

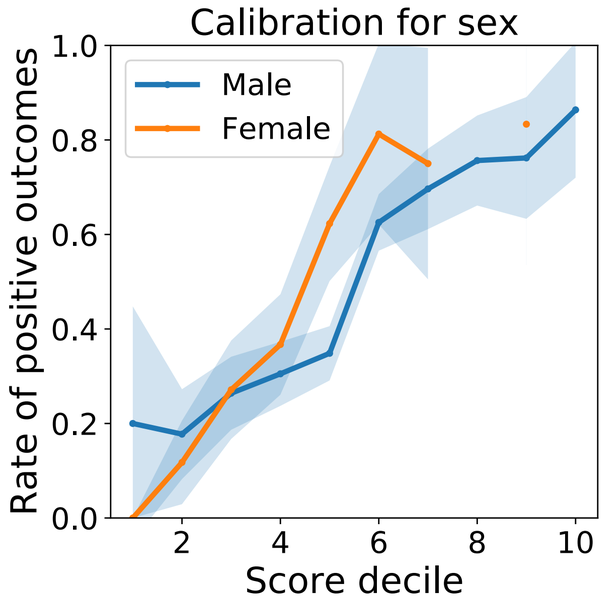




EXPERIMENTS ON FLORIDA PRETRIAL DEFENDANTS DATASET Angwin et al (2017)

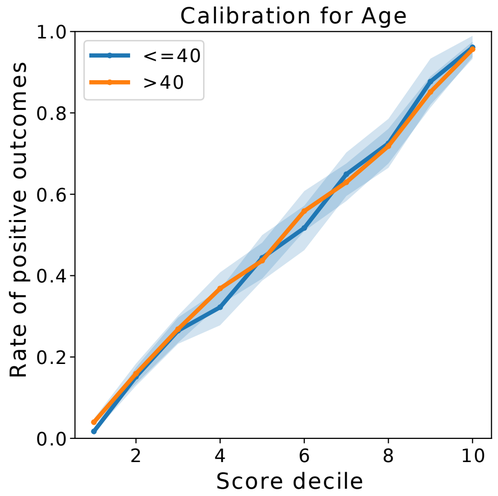
- 7 features, 7214 individuals, predict 2-year recidivism
- **Observation 1**: ERM score is close to **calibrated** by group even when trained without the group attribute. Or there is insufficient data to decide.

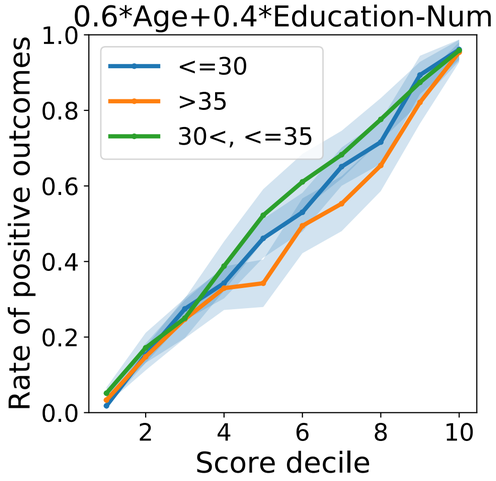


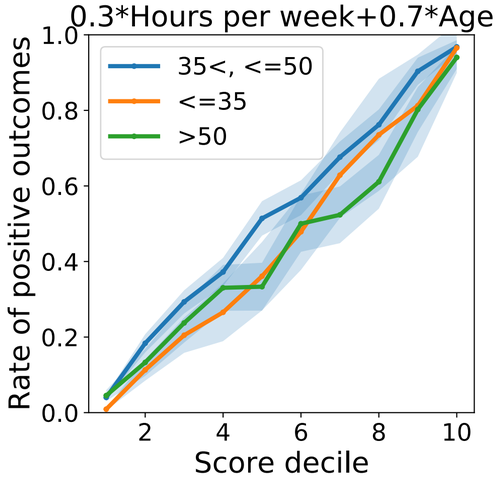


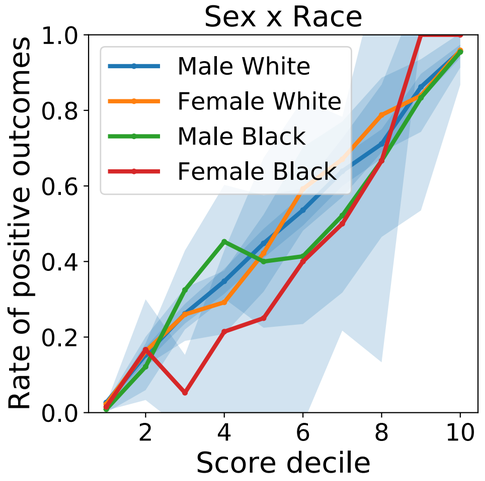
EXPERIMENTS ON UCI ADULT

- **Observation 2**: ERM score is simultaneously **calibrated** w.r.t. many group attributes including those defined post-hoc









TRADEOFFS WITH OTHER FAIRNESS CRITERIA

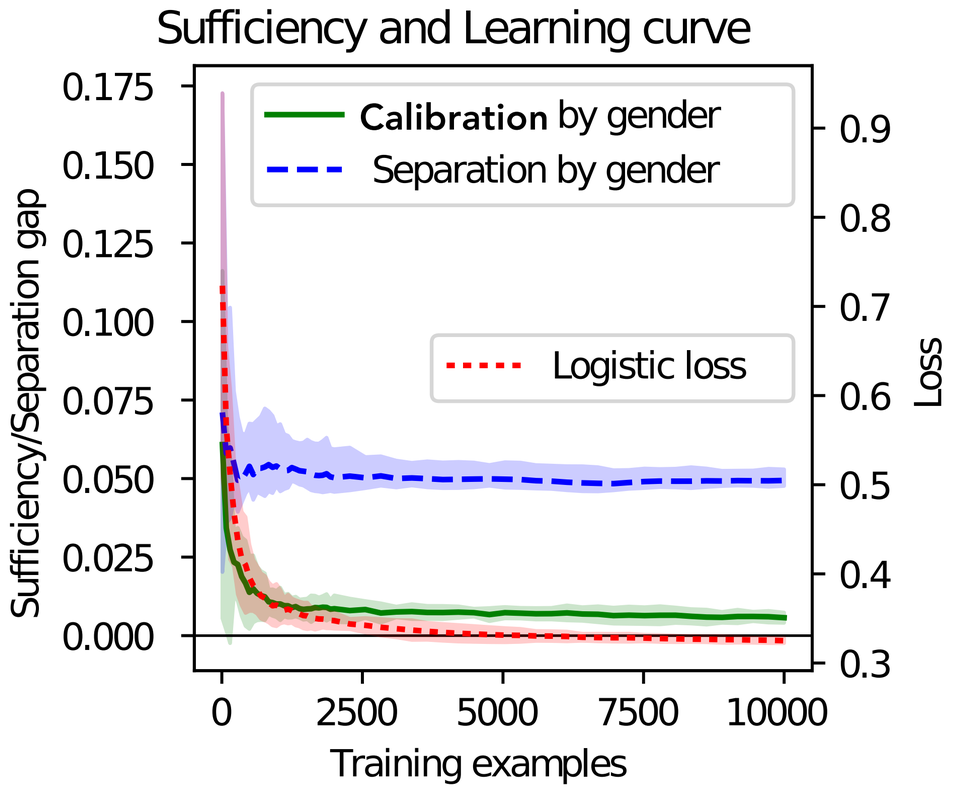
- Calibration has been suggested as a fairness criterion when A is a sensitive attribute [Kleinberg et al. 2016; Chouldechova, 2017].
- Other notions of group fairness include **separation** (aka equalized odds):

$$\mathbb{E}[f \mid A, Y] = \mathbb{E}[f \mid Y]$$

  - "Mean score for individuals with positive (negative) outcomes is same across groups"
- **Separation gap**: $\mathbf{sep}_f(A) := \mathbb{E}[\,|\mathbb{E}[f \mid A, Y] - \mathbb{E}[f \mid Y]|\,]$
- **Lower bound** (details in paper) shows that unconstrained ERM necessarily has a large **separation gap** that is problem-dependent

EXPERIMENTS ON UCI ADULT

- **Observation 3**: **Calibration** gap decreases to 0 as we increase the number of training examples. **Separation** gap does not.



LESSONS LEARNT

- Approximate group **calibration** is satisfied with unconstrained ERM without needing active intervention.
- Enforcing group **calibration** **does not require any departure from** unconstrained machine learning, which largely describes **current practice**.
- When should we be fine with group **calibration** as a normative fairness goal?
  - only if we're happy with unconstrained machine learning
  - harms of unconstrained machine learning [Crawford, 2013; Barocas and Selbst, 2016; Crawford, 2017]
- Practitioners hoping to deviate from current practice will not achieve this goal by asking for **calibration** alone.

THANK YOU

For more details and experiments, see full version:
https://arxiv.org/abs/1808.10013