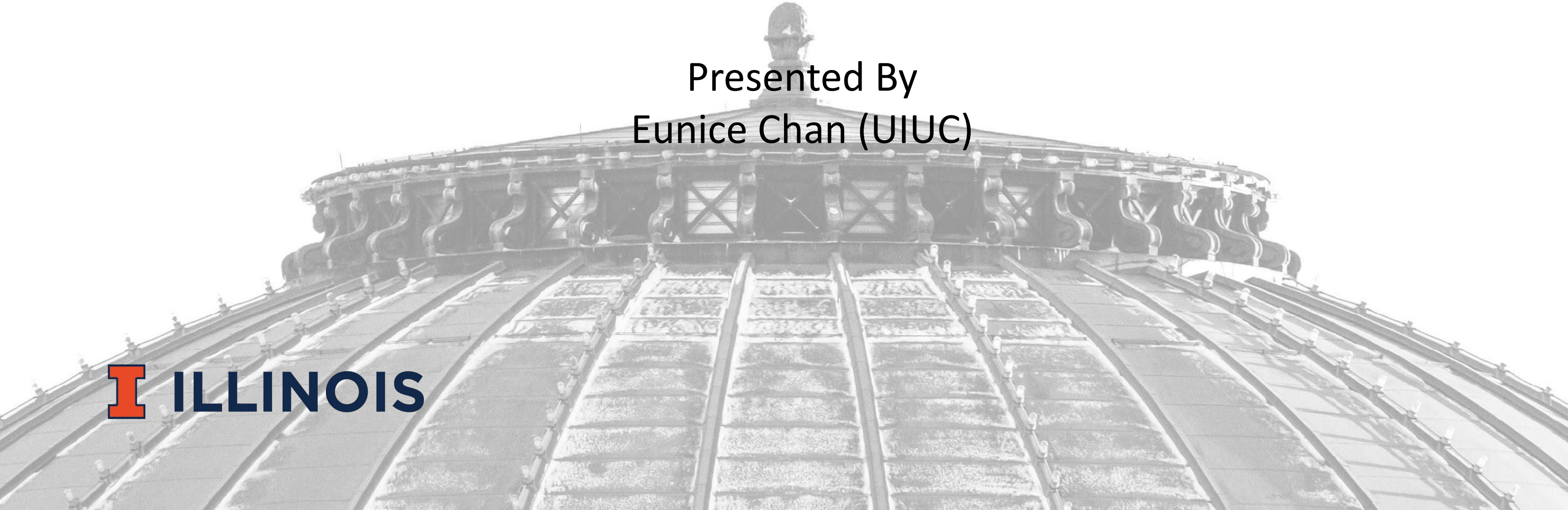
Fair Active Learning

Paper: https://arxiv.org/abs/2001.01796
Code: https://github.com/anahideh/FAL--Fair-Active-Learning

Paper By
Hadis Anahideh (UIC), Abolfazl Asudeh (UIC),
Saravanan Thirumuruganathan (QCRI, HBKU)

Presented By
Eunice Chan (UIUC)



ILLINOIS

Outline

1. Introduction
2. Toy Example
3. Proposed Approach
4. Experimental Results
5. Conclusions

Data-Driven Decision-Making in Court

- Criminal assessment algorithms
  - **Input**: Background information of individuals
  - **Output**: Recidivism scores
  - **Usage**: Setting bails or sentencing criminals.
- Discriminatory: assigns higher risks to African American individuals.

How to create fair machine learning algorithms?

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. ProPublica (2016). https://bit.ly/2s0UMfA

Defining Fairness

Demographic Parity: equal proportion of positive predictions in each group.

$$P(Y = 1|S = a) = P(Y = 1|S = b)$$

Y: Outcome

S: Sensitive attribute (e.g. race)

Defining the Problem

Models reflect the biases in their data

→ **Proposed Solution**

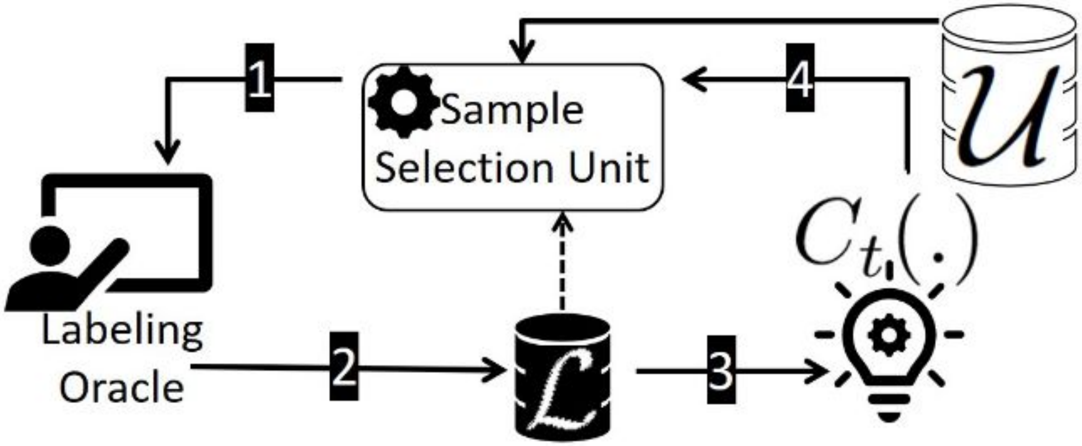
Try to build a dataset that will yield a fair(er) model

Active Learning Setting

Given

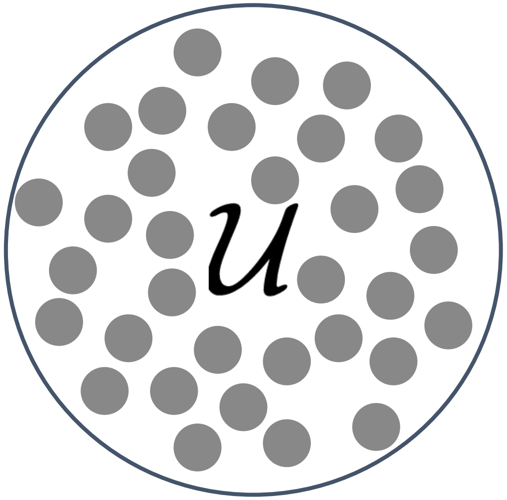
- A classifier $C: X \rightarrow Y$
- Pool of unlabeled data $\mathcal{U}$
  - Assumed i.i.d.

Active Learning (AL) identifies the data points to be labeled (placed in pool of labeled data $\mathcal{L}$ for training).



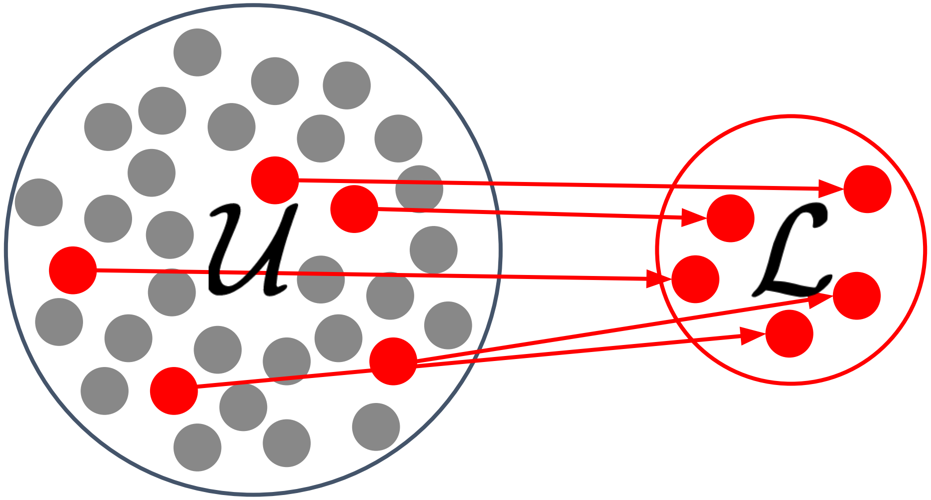
Fair Active Learning

Given a pool of unlabeled data, select a subset to be labeled and used to train a fair model.



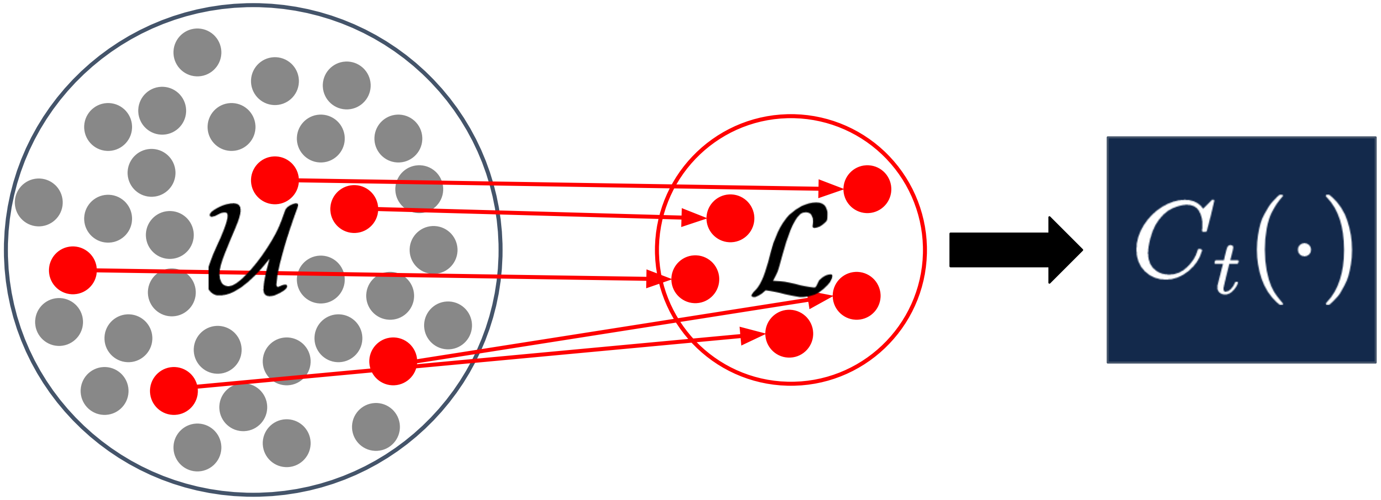
Fair Active Learning

Given a pool of unlabeled data, select a subset to be labeled and used to train a fair model.



Fair Active Learning

Given a pool of unlabeled data, select a subset to be labeled and used to train a fair model.



Outline

1. Introduction
2. Toy Example
3. Proposed Approach
4. Experimental Results
5. Conclusions

FAL by Covariance (FBC)

- Classifier is a generalized linear model

$$\hat{y} = \theta^\top X$$

- Lemma

$$cov(S, \hat{y}) = \sum_{i=1}^{d} \theta_i cov(S, x_i)$$

where S is the sensitive attribute (for fairness).

  - The covariance of the model with the sensitive attribute depends only on the weight vector θ and the underlying covariance of features X with S.

- Improving fairness by reducing the weight assigned to features highly correlated with the sensitive attribute ($cov(x_i, S)$ is high).

Generalization

Generalized Linear Classifier Case

Select data points that will reduce the unfairness in the model.

Select data points that will reduce the weight assigned to features highly correlated with the sensitive attribute.

↓

General Classifier Case

Select data points that will reduce the unfairness in the model.

How do we decide this?

Outline

1. Introduction
2. Toy Example
3. Proposed Approach
4. Experimental Results
5. Conclusions

Expected Unfairness Reduction

Select the point that is expected to impart the largest reduction to the current model unfairness, after acquiring its label.

We don’t know the label in advance. Instead, calculate the unfairness as an expectation over the possible labels.

$$E[\mathcal{F}_t^i] = \sum_{k=0}^{K-1} \mathcal{F}(C_t^{i,k})\mathbb{P}(y = k|X^{(i)})$$

The expected fairness of the model after adding point i is the sum of all changes in the fairness if point i has the label y=k weighted by the probability the label y=k given the input.

Incorporating Fairness in AL (FAL)

1. FAL α-aggregate (FAL-α)
2. FAL Nested (Nested)
3. FAL Nested-Append (N-App)

FAL α-aggregate

- Most straightforward way: Multi-objective optimization
  - “Regularizer”
  - Optimize for a balance between accuracy and fairness
- α can be fixed or change according to a decay function
- Drawback: How to determine the tradeoff?

$$\underset{\langle X^{(i)},S^{(i)}\rangle\in\mathcal{U}}{\operatorname{argmax}} \quad \alpha\boxed{\mathcal{H}_{t-1}(y^{(i)})} + (1-\alpha)\boxed{\left(\mathcal{F}(C_{t-1}) - \mathcal{F}(C_t^i)\right)}$$

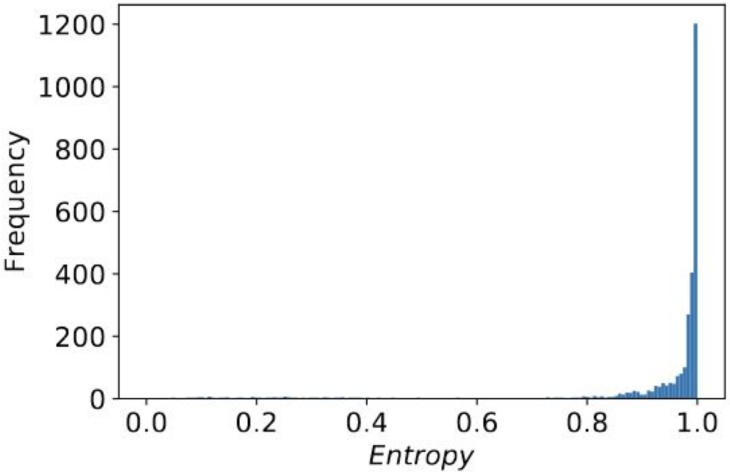
Shannon entropy
for misclassification error

change in demographic parity
for fairness

FAL Nested Motivation

Observation: In practice, the distribution of the entropy of the data points is right-skewed.

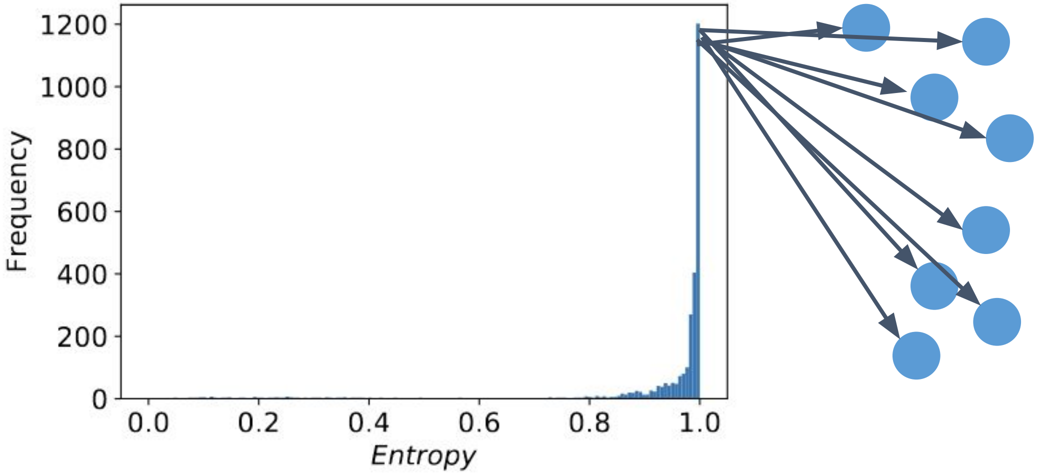
- Many points are almost equally good candidates from the accuracy perspective.
- We can reduce the computation cost of the model by only focusing on a subset of $\mathcal{U}$.



FAL Nested Algorithm

1. Select ℓ points that maximize entropy (for good accuracy).

$$\mathcal{U}_A = \underset{\langle X^{(i)}, S^{(i)} \rangle \in \mathcal{U}}{\ell\text{-argmax}} \; \mathcal{H}(y|X, \mathcal{L})$$

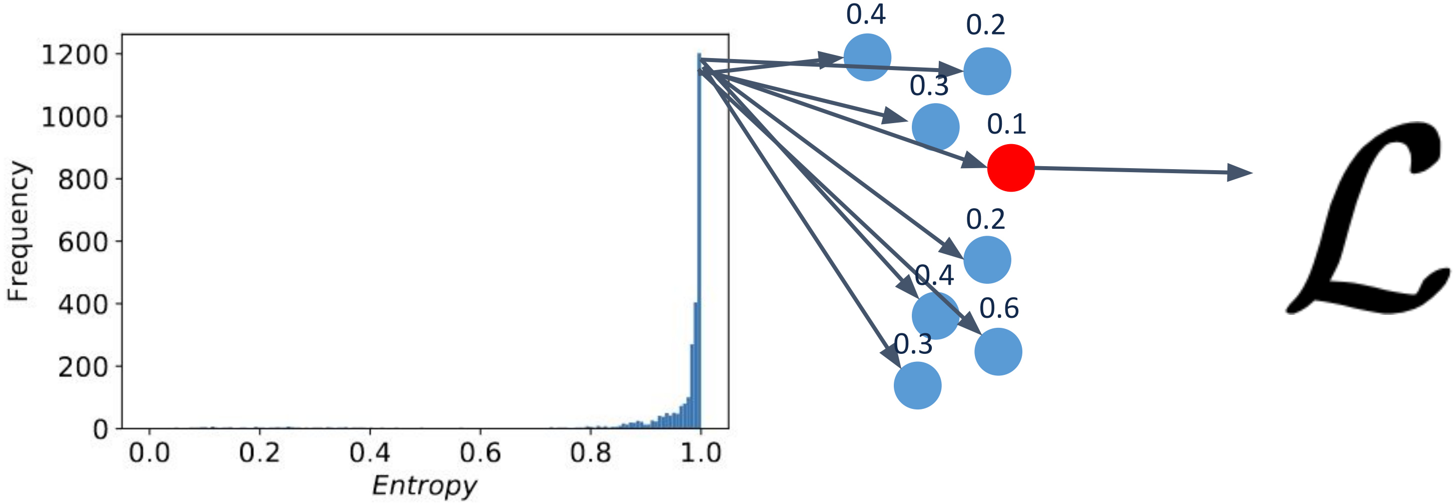


FAL Nested Algorithm

2. From those points, select the one that will maximizes the unfairness improvement.

$$\underset{\langle X^{(i)},S^{(i)}\rangle \in \mathcal{U}_A}{\operatorname{argmax}} \left(\mathcal{F}(C_{t-1}) - \mathcal{F}(C_t^i)\right)$$

$$= \underset{\langle X^{(i)},S^{(i)}\rangle \in \mathcal{U}_A}{\operatorname{argmin}} \mathcal{F}(C_t^i)$$



FAL Nested-Append

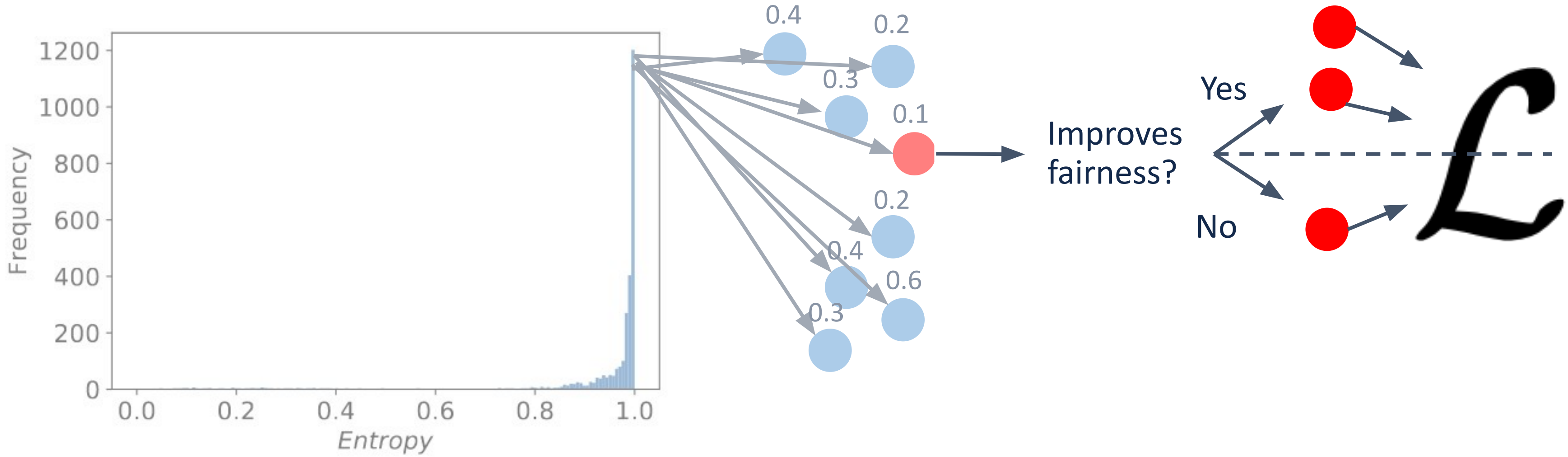
Observation: Expected unfairness reduction estimated by label likelihood. However, it could be wrong.

- A point may have a high expected unfairness reduction, but after being labeled, it increases unfairness in the model.

Solution: Modify FAL Nested.

If a point indeed reduces unfairness after being labeled, replicate the point in the labeled pool.

FAL Nested-Append



Outline

1. Introduction
2. Toy Example
3. Proposed Approach
4. Experimental Results
5. Conclusions

Datasets

COMPAS

- **Input**: background information on individual (age, race, marital status, prior convictions, charge degree, etc.)
- **Group**: race (black/white)
- **Output**: two-year recidivism (yes/no)

Adult

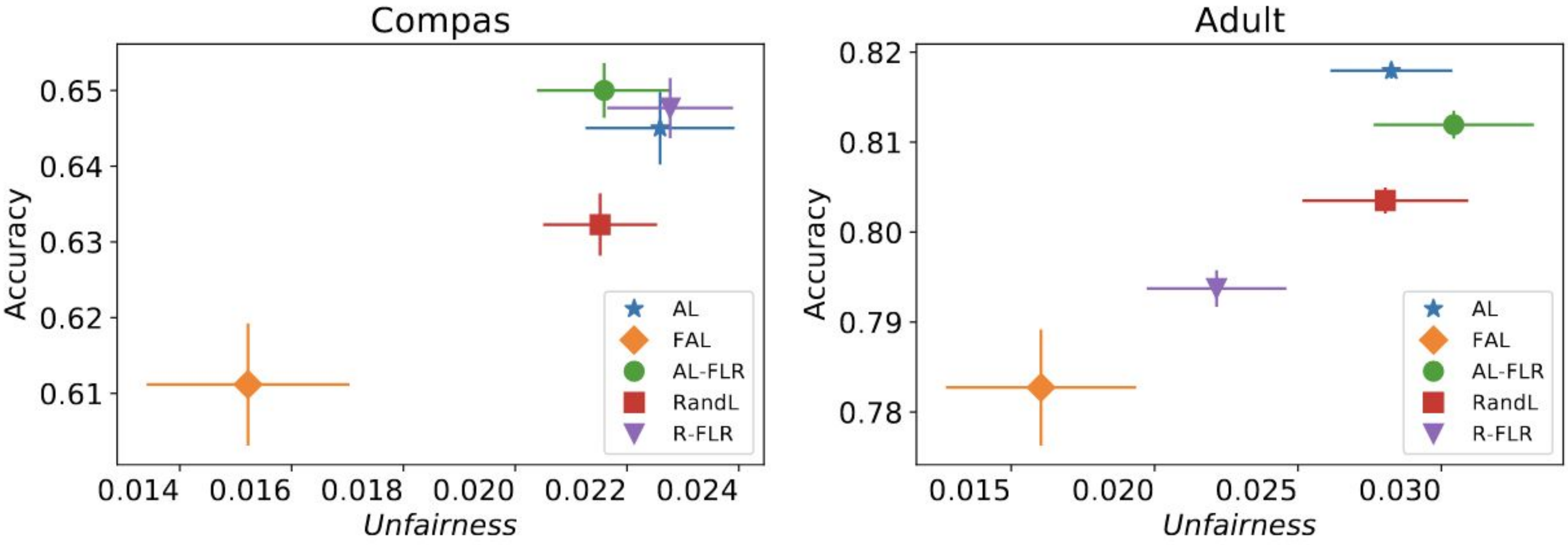
- **Input**: background information on individual (age, occupation, education, race, sex, marital status, hours-per-week, native country, etc.)
- **Group**: sex (male/female)
- **Output**: income (≥$50k yes/no)

Baselines

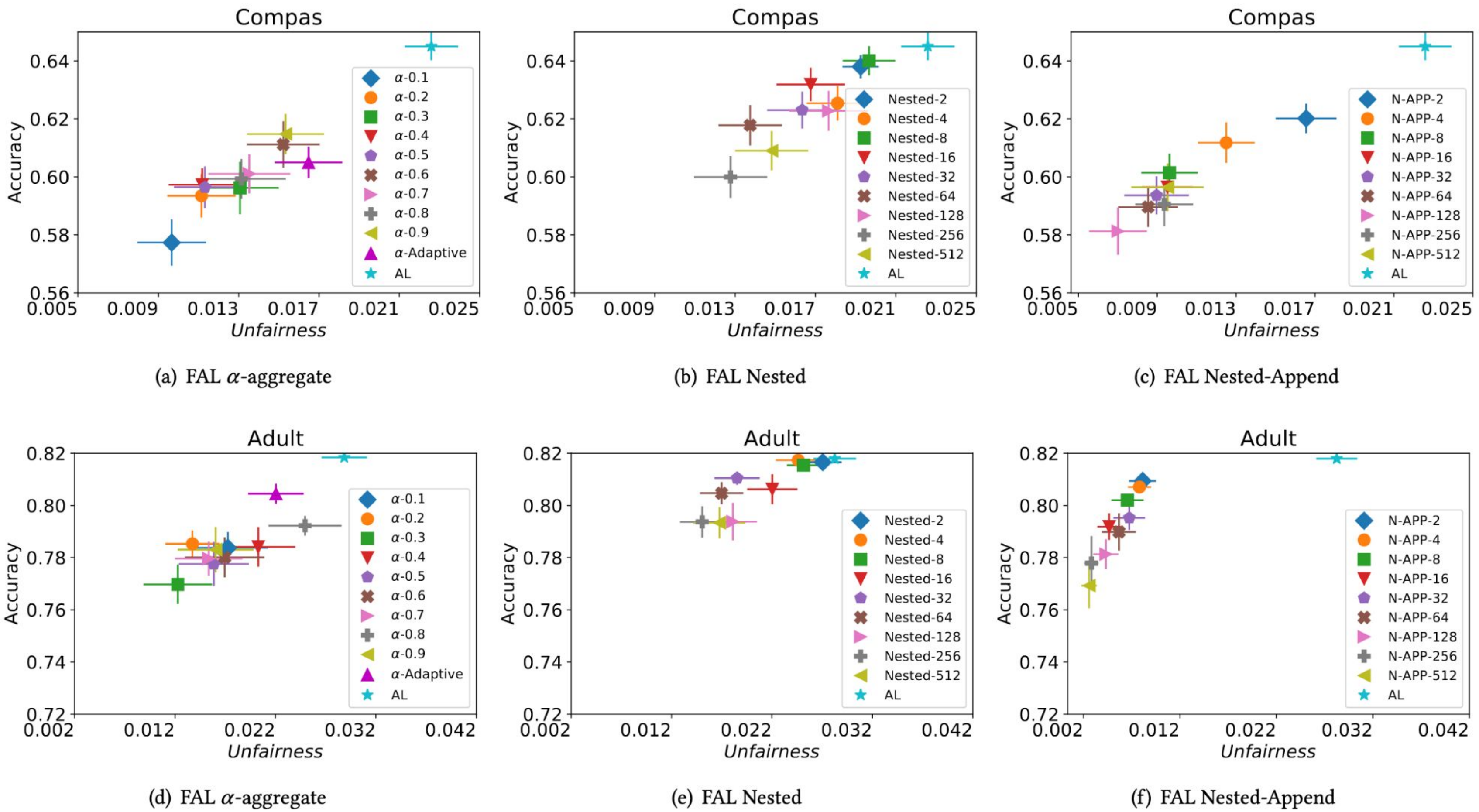
- **RandL**
  - Random sample
  - Fit logistic regression
- **R-FLR**
  - Random sample
  - Fit fair logistic regression
- **AL**
  - Sample based on informativeness
  - Fit logistic regression
- **AL-FLR**
  - Sample based on informativeness
  - Fit fair logistic regression

FAL-α vs Baselines

- Compare FAL-α with α=0.6 against baselines.
- FAL-α significantly reduces unfairness while sustaining a comparable accuracy.



Comparing FAL Performance



(a) FAL α-aggregate

(b) FAL Nested

(c) FAL Nested-Append

(d) FAL α-aggregate

(e) FAL Nested

(f) FAL Nested-Append

Outline

1. Introduction
2. Toy Example
3. Proposed Approach
4. Experimental Results
5. Conclusions

Conclusions

Problem: Incorporating fairness into active learning

Solution:

- FAL α-aggregate (FAL-α)
- FAL Nested (Nested)
- FAL Nested-Append (N-App)
- FAL by Covariance (FBC)

Results:

- Produces a fairer model without significantly sacrificing the accuracy.
- Discuss variations for improving accuracy, fairness, and running time.

Weaknesses

- Although the fairer model does not significantly sacrifice the accuracy, the accuracy is reduced up to 6% depending on tuning.
- Only deals with selecting 1 point per round.
- Slow because calculating the unfairness reduction requires training a version of the model on the hypothetical case point = label for each label.