Fair Multiple Decision Making Through Soft Interventions

Yaowei Hu[1], Yongkai Wu[2], Lu Zhang[1], Xintao Wu[1]

[1]University of Arkansas
[2]Clemson University

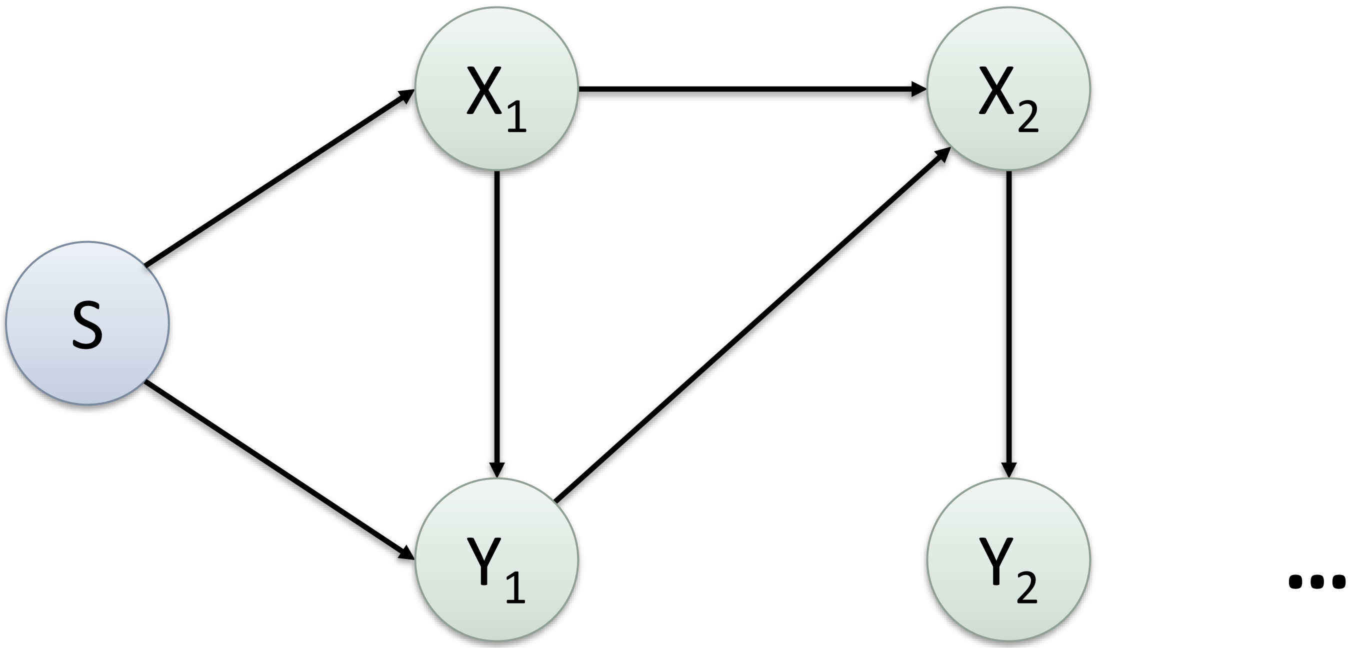
Fair Decision Making

- **Admission**
- **Recruiting**
- **Credit**

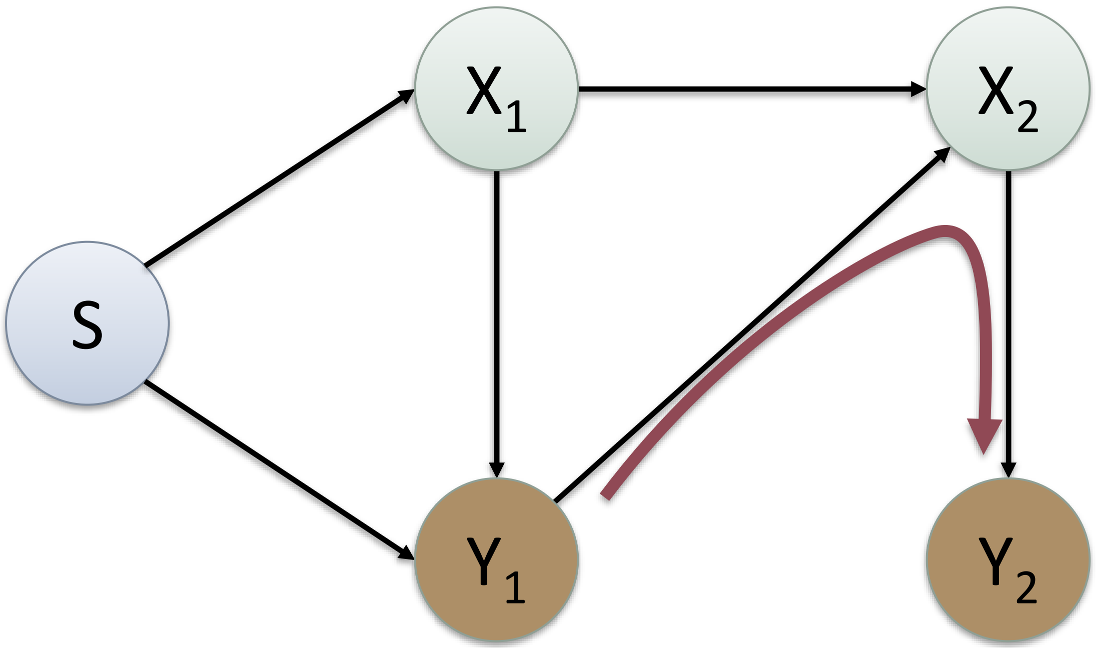
…

How to ensure fairness in algorithmic decision making models is an important task in machine learning.

Multiple Decision Making



Fair Multiple Decision Making



What if we build fair model for each task independently?

$(X_1, Y_1) \quad (X_2, Y_2)$

Step 1: data collection

$(X_1, Y_1) \Longrightarrow h_1$ **(fair classifier)**

$(X_2, Y_2) \Longrightarrow h_2$ **(fair classifier)**

Step 2: offline training and evaluation (separately)

$\hat{X}_1 \xrightarrow{h_1} \hat{Y}_1$ **(fair)**

$\hat{X}_2 \xrightarrow{h_2} \hat{Y}_2$ **(unfair)**

Step 3: deploy and make decisions on new data

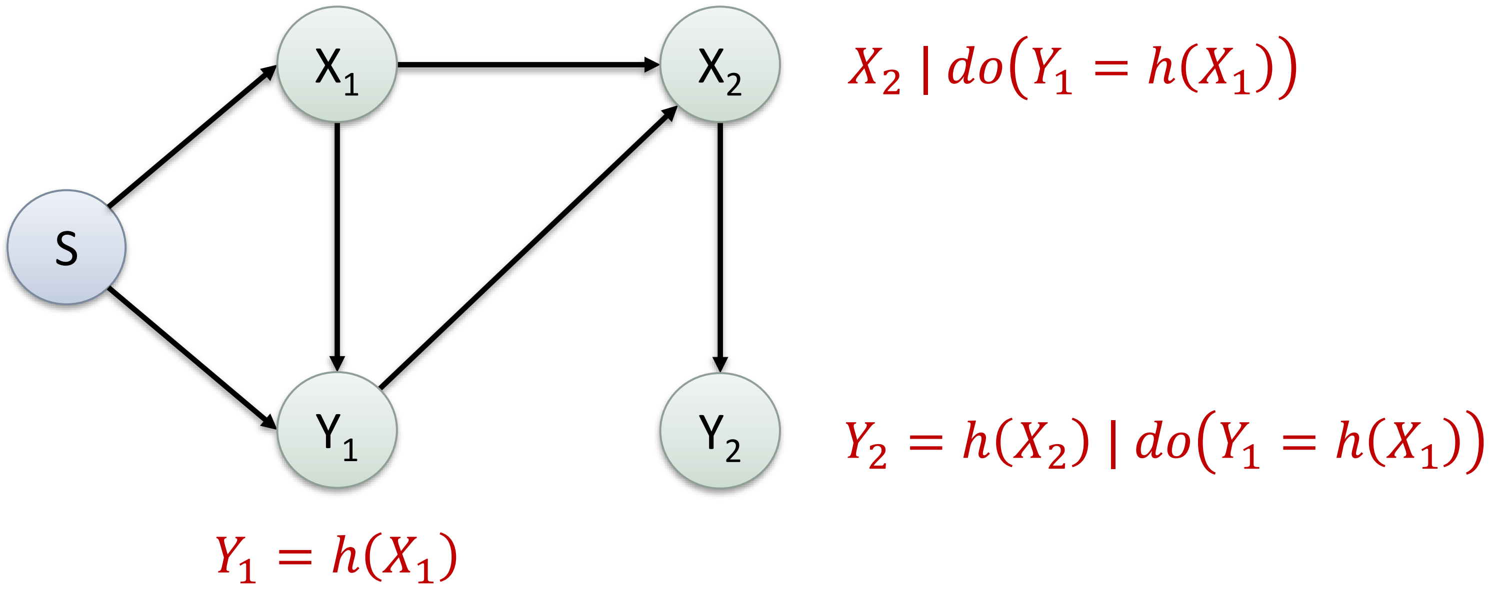
Why ?

- **Decision $\widehat{Y}_1$ will affect values of $\widehat{X}_2$**
- **Distribution $X_2 \neq$ Distribution $\widehat{X}_2$**

Proposed Solution

<u>Core idea:</u> leverage Pearl's structural causal model (SCM), treat each decision model as a soft intervention and infer the post-intervention distributions to formulate the loss function as well as the fairness constraints.

Using Soft Interventions to Simulate Decision Model Deployments



- In general, we have l decisions $\{Y_1, \cdots, Y_l\}$.
- For each decision Y_k, we build a classifier $h_k(\boldsymbol{z}_k)$.
- The soft intervention for deploying all these models is $do(h_1, \cdots, h_l)$.

Loss Function and Fair Constraints

- Traditionally, classification error of classifier $h: \boldsymbol{Z} \mapsto Y$ is

$$R(h) = \mathbb{E}_{\boldsymbol{Z}}\left[P(Y = 1|\boldsymbol{z})\mathbf{1}_{h(\boldsymbol{z})<0} + P(Y = 0|\boldsymbol{z})\mathbf{1}_{h(\boldsymbol{z})\geq 0}\right]$$

- Under soft intervention of deploying all models, for classifier h_k

$$R(h_k) = \mathbb{E}_{\boldsymbol{Z}_k|do(h_1,\cdots,h_l)}\left[P(Y = 1|\boldsymbol{z}_k)\mathbf{1}_{h(\boldsymbol{z}_k)<0} + P(Y = 0|\boldsymbol{z}_k)\mathbf{1}_{h(\boldsymbol{z}_k)\geq 0}\right]$$

- Similarly, fairness constraints are given by total effect

$$T(h_k) = P(Y = 1|do(S = 1, h_1, \cdots, h_l)) - P(Y = 1|do(S = 0, h_1, \cdots, h_l))$$

Deriving Loss Function and Fair Constraints with Observed Data

- Loss function

$$
\begin{aligned}
R_\phi(h_k) = \mathop{\mathbb{E}}_{S,\mathbf{X}'_{Y_k}} \Bigg[& P(y_k^+|\mathbf{z}_k)\phi(h_k(\mathbf{z}_k)) \sum_{\mathbf{Y}'_{Y_k}} \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+} \phi(-h_i(\mathbf{z}_i)) \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-} \phi(h_i(\mathbf{z}_i)) \prod_{X_i\in\mathbf{X}'_{Y_k}} \frac{P(\mathbf{y}'_{X_i}|s,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s,\mathbf{x}'_{X_i})} \\
& + P(y_k^-|\mathbf{z}_k)\phi(-h_k(\mathbf{z}_k)) \sum_{\mathbf{Y}'_{Y_k}} \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+} \phi(-h_i(\mathbf{z}_i)) \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-} \phi(h_i(\mathbf{z}_i)) \prod_{X_i\in\mathbf{X}'_{Y_k}} \frac{P(\mathbf{y}'_{X_i}|s,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s,\mathbf{x}'_{X_i})} \Bigg].
\end{aligned}
$$

- Fair constraint

$$
\begin{aligned}
T_\phi(h_k) = & \mathop{\mathbb{E}}_{\mathbf{X}'_{Y_k}|S=s^+} \Bigg[\phi(-h_k(\mathbf{z}_k)) \sum_{\mathbf{Y}'_{Y_k}} \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+} \phi(-h_i(\mathbf{z}_i)) \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-} \phi(h_i(\mathbf{z}_i)) \prod_{X_i\in\mathbf{X}} \frac{P(\mathbf{y}'_{X_i}|s^+,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s^+,\mathbf{x}'_{X_i})} \Bigg] \\
& + \mathop{\mathbb{E}}_{\mathbf{X}'_{Y_k}|S=s^-} \Bigg[\phi(h_k(\mathbf{z}_k)) \sum_{\mathbf{Y}'_{Y_k}} \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+} \phi(-h_i(\mathbf{z}_i)) \prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-} \phi(h_i(\mathbf{z}_i)) \prod_{X_i\in\mathbf{X}} \frac{P(\mathbf{y}'_{X_i}|s^-,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s^-,\mathbf{x}'_{X_i})} \Bigg] - 1.
\end{aligned}
$$

Problem Formulation for Fair Multiple Decision Making

- *The problem of fair multiple decision making for $\boldsymbol{Y} = \{Y_1, \cdots, Y_l\}$ is formulated as the following constrained optimization problem:*

$$\min_{h_1,\cdots,h_l \in \mathcal{H}} \sum_{k=1}^{l} R_\phi(h_k) \qquad s.t. \qquad \forall k, -\tau_k \leq T_\phi(h_k) \leq \tau_k$$

where $R_\phi(h_k)$ and $T_\phi(h_k)$ are smoothed loss function and fair constraint.

Excess Risk Bound

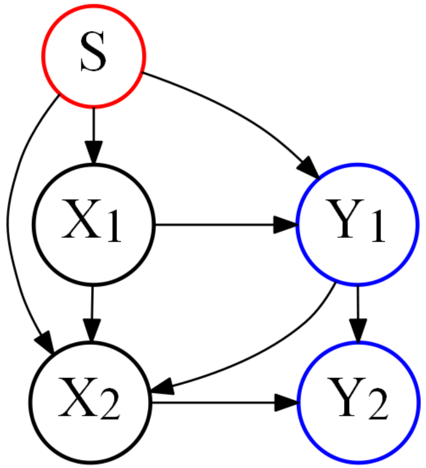
- *For any classification-calibrated surrogate function* ϕ *satisfying* $\phi(0) = 1$ *and* $\inf_{\alpha \in \mathbb{R}} \phi(\alpha) = 0$, *any measurable function* h_k *for predicting* Y_k, *we have*

$$\psi(R(h_k) - R^*) \leq R_\phi(h_k) - R_\phi^*$$

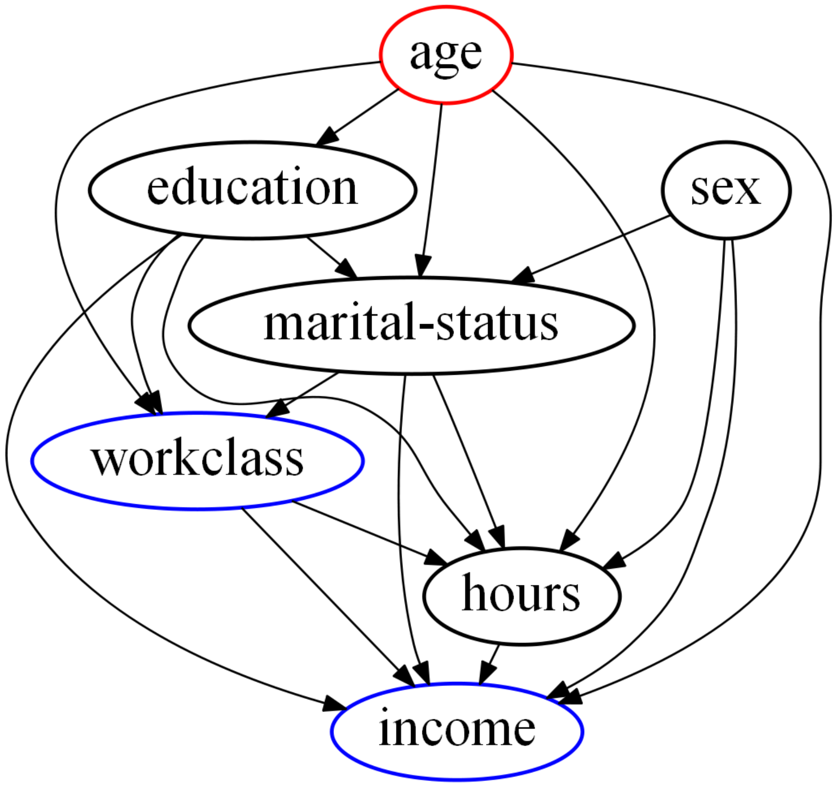
where ψ *is a non-decreasing function mapping from* $[0,1]$ *to* $[0,\infty)$.

Experiments

- Data:
  - Synthetic data:



  - Adult data:



Experiments

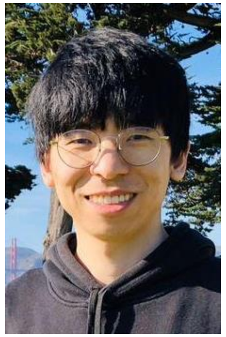
Table 1: Accuracy and unfairness from Unconstrained, Separate, Serial and Joint methods on synthetic and Adult data (bold values indicate violation of fairness).

Phase			Synthetic				Adult			
			Uncons.	Separate	Serial	Joint	Uncons.	Separate	Serial	Joint
Train	h_1	Acc. (%)	80.32	75.35	75.35	75.35	55.71	55.64	55.63	55.63
		Unfairness	**0.15**	0.01	0.01	0.01	**0.15**	0.05	0.05	0.05
	h_2	Acc. (%)	90.13	75.79	84.02	82.77	76.75	71.17	68.90	69.31
		Unfairness	**0.23**	0.04	0.03	0.04	**0.24**	0.10	0.10	0.10
Test	h_1	Acc. (%)	80.70	75.54	75.54	75.54	55.63	55.56	55.57	55.57
		Unfairness	**0.15**	0.01	0.01	0.01	**0.15**	0.05	0.05	0.05
	h_2	Acc. (%)	89.95	77.06	84.16	82.09	77.07	73.33	68.91	69.40
		Unfairness	**0.13**	**0.09**	0.03	0.03	**0.23**	**0.17**	0.10	0.10

Conclusions

- Proposed an approach that learns multiple fair classifiers from a static training dataset.
- Treated the deployment of each classifier as a soft intervention and inferred the distributions after the deployment as post-intervention distributions.
- Adopted surrogate functions to smooth the loss function and fair constraints to formulate the fair classification problem as a constrained optimization problem.
- Theoretically analyzed excess risk bound.
- Conducted experiments on both synthetic and real-world datasets.

Fair Multiple Decision Making Through Soft Interventions




Yaowei Hu **Yongkai Wu** **Lu Zhang** **Xintao Wu**

This work was supported in part by NSF 1646654, 1920920, and 1946391.