UNIVERSITY OF MARYLAND

CLEMSON UNIVERSITY

TOWARDS COUNTERFACTUAL FAIRNESS THROUGH AUXILIARY VARIABLES

Bowei Tian[1], Ziyao Wang[1], Shwai He[1], Wanghao Ye[1]
Guoheng Sun[1], Yucong Dai[2], Yongkai Wu[2*], Ang Li[1*]
[1]University of Maryland, College Park, [2]Clemson University
{btian1, ziyaow, shwaihe, wy891, ghsun, angliece}@umd.edu
yucongd@g.clemson.edu, yongkaw@clemson.edu

1 INTRODUCTION

We observe that the majority of existing methods assume sensitive attributes should not be causally influenced by any other variables (Kusner et al., 2017; Berk et al., 2021; Ma et al., 2023). This assumption overlooks the essentials of sensitive features, i.e., which part of the sensitive feature is intrinsic or essential for the inference and which part should be neglected. Also, existing methods usually fit into a specific scenario where the causal relationship from the sensitive attribute to the target attribute is fixed. For example, race should generally not influence decision-making at all, making it hard to extend and distribute in real-world scenarios. For example, in a demography experiment, race distribution can be deduced from the geographic distribution of the population, which can not be causally neglected.

To tackle these challenges, we propose a novel framework, EXOC, which introduces intuitive modifications to the causal model. This framework utilizes the auxiliary variables in causal inference, extracting essential information from sensitive attributes and effectively controlling the flow of information from the sensitive attribute to the target attribute. We summarize our contributions as follows:

- We develop a framework that utilizes the auxiliary variables in causal inference, extracting essential information from sensitive attributes and enhancing fairness without sacrificing much accuracy.
- We formalize a method to regulate the flow of information from the sensitive attribute to the target attribute, effectively controlling the balance between accuracy and fairness.
- We provide theoretical analysis and conduct extensive baseline and ablation experiments to validate the effectiveness of our approach.

2 PRELIMINARIES

2.1 COUNTERFACTUAL FAIRNESS

Counterfactual fairness: Given a predictor and a factual condition $\mathbf{O} = \mathbf{o}$, the predictor makes the prediction for each instance. The predictor is *counterfactually fair* Kusner et al. (2017); Wu et al. (2019) if under any context $\mathbf{o}$,

$$P\left(Y_{S \leftarrow s} = y \mid \mathbf{o}\right) = P\left(Y_{S \leftarrow s'} = y \mid \mathbf{o}\right), \tag{1}$$

for all y and counterfactual $s' \neq s$, where $\mathbf{O} = \{S, \mathbf{X}\}$, S is the sensitive feature and $\mathbf{X}$ is all remaining relative features. Here $P\left(Y_{S \leftarrow s} = y \mid \mathbf{o}\right)$ denotes on condition of $\mathbf{O}$, the prediction made on the counterfactuals when the value of S *had been* set to s.

Approximate counterfactual fairness: A predictor $Y = f(\mathbf{O})$ satisfies $(\varepsilon, 0)$-*approximate counterfactual fairness* Russell et al. (2017) if, given the factual condition $\mathbf{O} = \mathbf{o}$, we have:

$$|[(Y_{S \leftarrow s} - Y_{S \leftarrow s'}) \mid \mathbf{o}]| \leq \varepsilon \tag{2}$$

for all counterfactual $s' \neq s$. This metric is extra useful for measuring counterfactual fairness in individual manners. We refer to this metric as counterfactual fairness in our theoretical analysis unless otherwise specified.

3 Design

3.1 Causal model overview

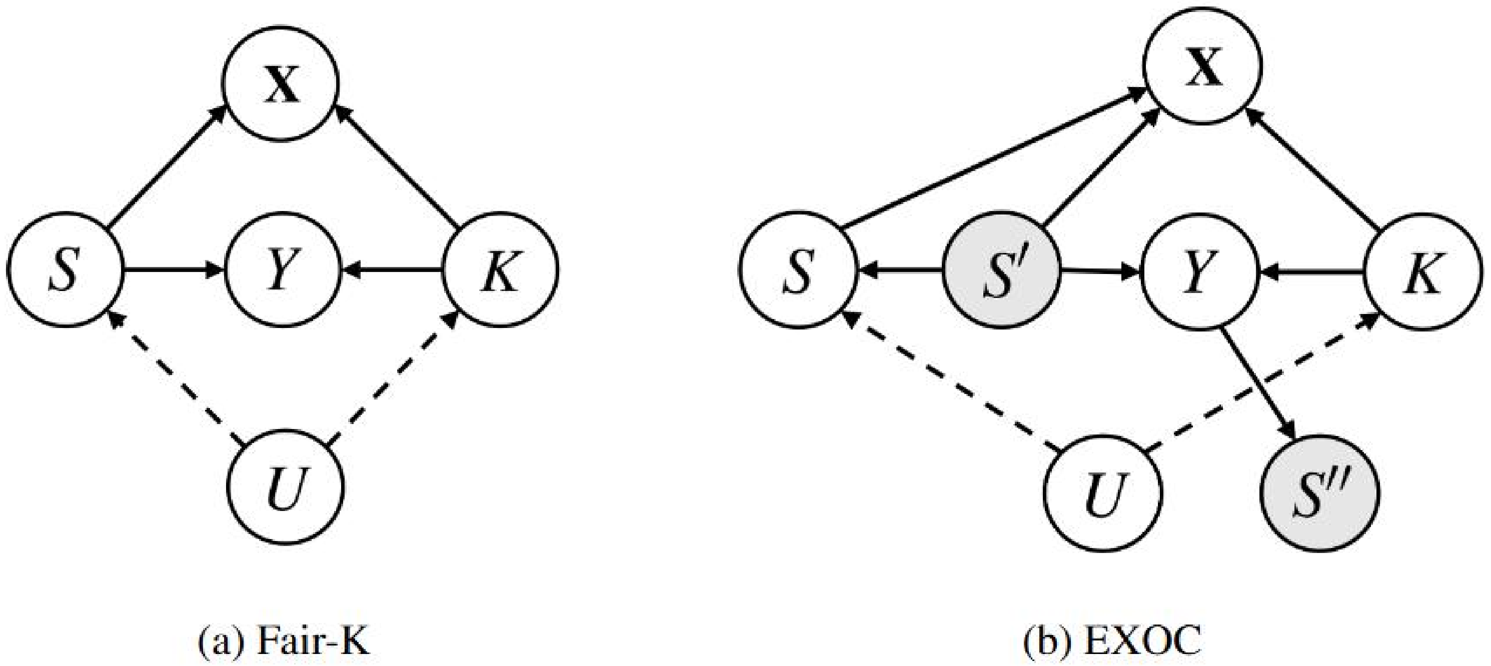


Figure 1: The causal models of Fair-K and EXOC. S is the sensitive attribute, $\mathbf{X}$ is observed non-sensitive attributes, Y is the target attribute, K is the latent domain knowledge, and S' and S'' are latent auxiliary nodes, U is the exogenous variable. The solid lines represent designed causal relationships, and dashed lines mean our focused existing relationships in implementation, illustrated in 3.2.2 (note that U have existing causal relationships with every node (Pearl, 2009)).

3.2 Illustration of S': substitution node

3.2.1 Causal inference

In Fig. 1a, for each *individual*, the causal relationship to Y can be written as:

$$Y = \alpha S + \beta K,$$

Similar in Fig. 1b, we have:

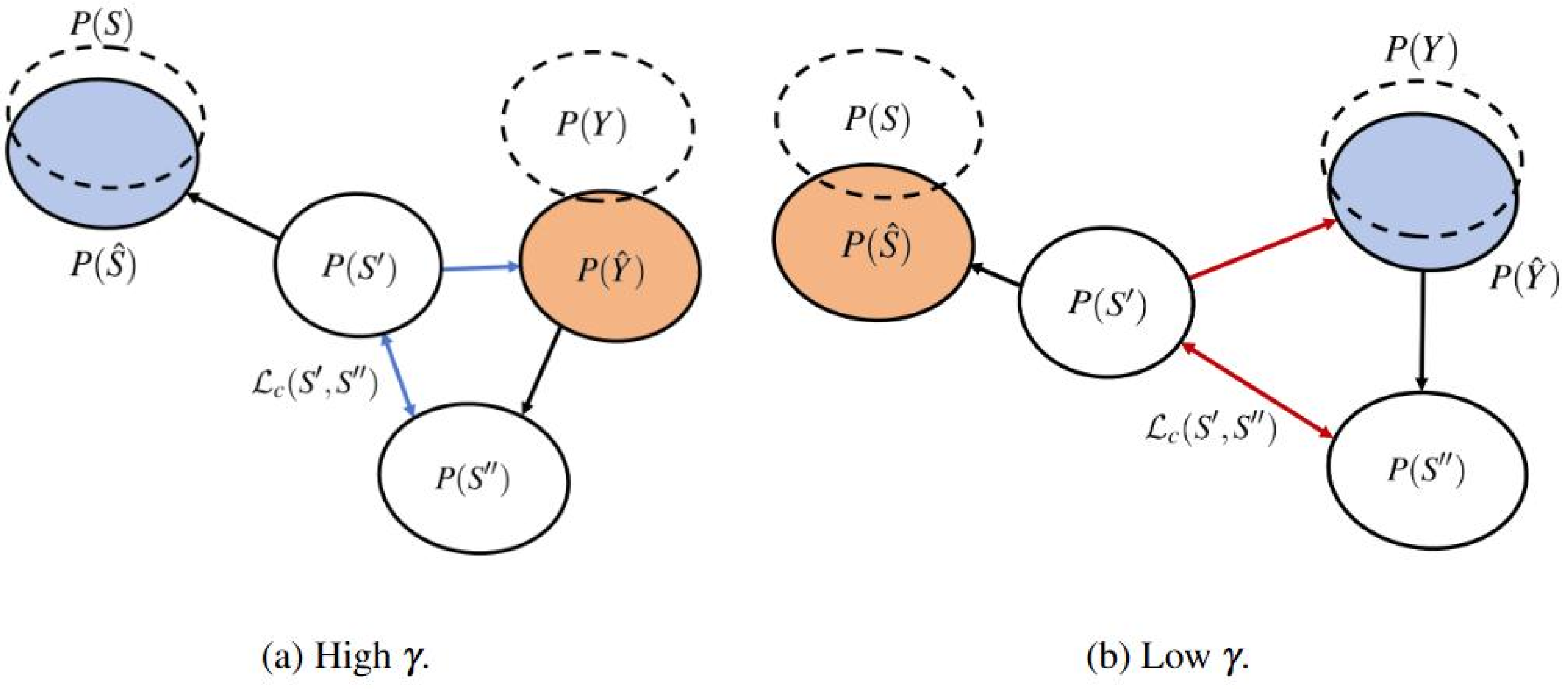
$$Y = \tilde{\alpha} S' + \tilde{\beta} K,$$

approximate counterfactual fairness bound in these equations as:

$$\begin{aligned} |[Y_{S \leftarrow s^*} - Y_{S \leftarrow s} \mid \mathbf{o}]_a| &\leq \left| \alpha(s^* - s) \pm 3\sqrt{2} \cdot |\beta| \sigma_K \right| \\ &= |\alpha(s^* - s)| + 3\sqrt{2} \cdot |\beta| \sigma_K = \varepsilon_a, \end{aligned} \tag{13}$$

$$|[Y_{S \leftarrow s^*} - Y_{S \leftarrow s} \mid \mathbf{o}]_b| \leq 3\sqrt{2\left(\tilde{\alpha}^2 \tilde{\sigma}_{S'}^2 + \tilde{\beta}^2 \tilde{\sigma}_K^2\right)} = \varepsilon_b, \tag{14}$$

3.3 ILLUSTRATION OF S'': CONTROL NODE



(a) High γ.

(b) Low γ.

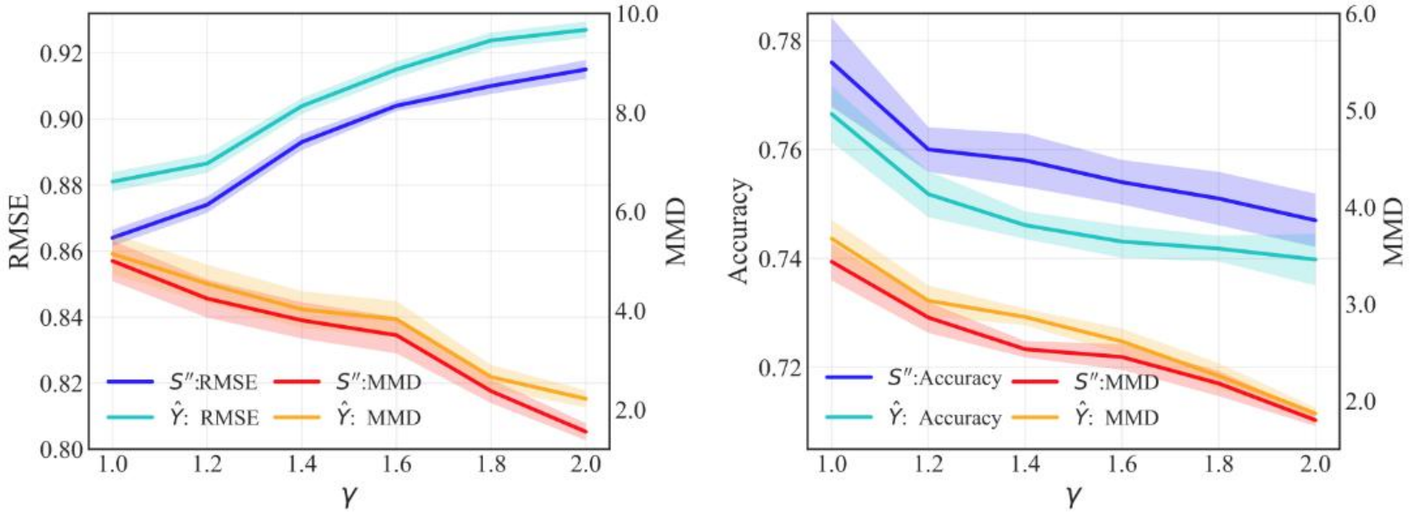
Figure 2: The distribution mapping regarding $\mathcal{L}_c(S',S'')$, which can be seen as a probability inference perspective of the partial causal graph in Fig. 1b, where blue arrows mean the distribution parity are tightened, red arrows mean loosened, the dashed line circle is the true distribution, the full line circle is the inferred distribution. Blue circles are close to true distributions, and orange circles are far from true distributions. Note that the γ is positively related to the effect of $\mathcal{L}_c(S',S'')$, so the constraint of $\mathcal{L}_c(S',S'')$ in Fig 2a is tighter, in Fig 2b is looser.

4 EXPERIMENTS

Table 1: The comparison on synthetic datasets among Constant, Full, Unaware, Fair-K (Kusner et al., 2017), CLAIRE (Ma et al., 2023) and EXOC (Ours) on Law school (Krueger et al., 2021) and Adult (Becker & Kohavi, 1996) dataset.

Method	*Law school*				*Adult*		
	RMSE (↓)	MAE (↓)	MMD (↓)	Wass (↓)	Accuracy (↑)	MMD (↓)	Wass (↓)
Constant	$0.938_{\pm 0.004}$	$0.759_{\pm 0.006}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.737_{\pm 0.006}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$
Full	$0.862_{\pm 0.005}$	$0.689_{\pm 0.005}$	$278.918_{\pm 25.814}$	$69.248_{\pm 6.136}$	$0.807_{\pm 0.005}$	$52.515_{\pm 3.757}$	$6.116_{\pm 0.637}$
Unaware	$0.900_{\pm 0.008}$	$0.726_{\pm 0.007}$	$40.256_{\pm 3.187}$	$10.256_{\pm 1.187}$	$0.804_{\pm 0.008}$	$19.732_{\pm 2.480}$	$2.004_{\pm 0.478}$
Fair-K	$0.894_{\pm 0.006}$	$0.718_{\pm 0.006}$	$4.313_{\pm 0.393}$	$3.733_{\pm 0.267}$	$0.745_{\pm 0.002}$	$3.597_{\pm 0.256}$	$1.553_{\pm 0.173}$
CLAIRE	$0.897_{\pm 0.002}$	$0.719_{\pm 0.002}$	$6.717_{\pm 0.492}$	$4.073_{\pm 0.139}$	$0.748_{\pm 0.005}$	$4.760_{\pm 0.275}$	$1.584_{\pm 0.203}$
EXOC	$0.874_{\pm 0.003}$	$0.702_{\pm 0.003}$	$3.824_{\pm 0.553}$	$3.590_{\pm 0.259}$	$0.760_{\pm 0.005}$	$2.958_{\pm 0.124}$	$1.428_{\pm 0.095}$

4.3 Ablation Study



(a) *Law school*

(b) *Adult*

Figure 3: The ablation study on S'' and $\hat{Y}$

Table 3: The ablation study on γ.

γ	*Law school*				*Adult*		
	RMSE (↓)	MAE (↓)	MMD (↓)	Wass (↓)	Accuracy (↑)	MMD(↓)	Wass (↓)
1	$0.875_{\pm 0.006}$	$0.698_{\pm 0.006}$	$4.489_{\pm 0.571}$	$3.905_{\pm 0.305}$	$0.765_{\pm 0.006}$	$3.532_{\pm 0.283}$	$1.539_{\pm 0.245}$
1.2	$0.867_{\pm 0.003}$	$0.706_{\pm 0.003}$	$3.832_{\pm 0.623}$	$3.580_{\pm 0.256}$	$0.760_{\pm 0.005}$	$2.961_{\pm 0.124}$	$1.426_{\pm 0.095}$
1.4	$0.886_{\pm 0.003}$	$0.724_{\pm 0.005}$	$3.377_{\pm 0.452}$	$3.352_{\pm 0.253}$	$0.755_{\pm 0.006}$	$2.628_{\pm 0.107}$	$1.352_{\pm 0.089}$
1.6	$0.900_{\pm 0.002}$	$0.728_{\pm 0.005}$	$3.089_{\pm 0.421}$	$3.203_{\pm 0.251}$	$0.757_{\pm 0.005}$	$2.458_{\pm 0.109}$	$1.297_{\pm 0.098}$
1.8	$0.903_{\pm 0.003}$	$0.731_{\pm 0.005}$	$2.034_{\pm 0.322}$	$2.890_{\pm 0.241}$	$0.751_{\pm 0.006}$	$2.068_{\pm 0.085}$	$1.204_{\pm 0.089}$
2	$0.909_{\pm 0.003}$	$0.735_{\pm 0.004}$	$1.342_{\pm 0.121}$	$2.824_{\pm 0.204}$	$0.746_{\pm 0.006}$	$1.792_{\pm 0.074}$	$1.184_{\pm 0.069}$

5 CONCLUSION

- We propose EXOC, a novel framework that enhances counterfactual fairness by introducing auxiliary variables to reveal and control intrinsic information flow from sensitive attributes. EXOC not only improves fairness but also maintains competitive accuracy, as validated through extensive experiments on synthetic and real-world datasets. Future directions include scaling to complex data, optimizing implementation efficiency, and further exploring the interplay between fairness, causality, and probabilistic modeling.