Non-Markovian Incentivized Counterfactual Fairness

As widespread adoption of automated decision-making progresses, so does the need to ensure that the decisions made are fair: namely, that the decisions are not influenced by preselected sensitive factors. Counterfactual fairness is a causal-aware notion of fairness, which compares the realized outcome for an individual to the counterfactual outcome where a sensitive attribute is changed, but all else stays the same. Furthermore, recent research has revealed graphical criteria which identifies which attributes an optimal algorithm will be incentivized to be unfair towards, in an unconfounded, Markovian setting.

Here, we extend the identification of incentivised counterfactual fairness to Non-Markovian (confounded) settings. We provide an efficient algorithm for this identification, and demonstrate that it is sound and complete.

Causal Inference under Optimization Pressure

Create a new type of intervention “O_R(x)”, which optimizes X based on the utility node R. How does this affect the distributions of other variables? Can we derive Goodhart’s law rigorously?

Bounding Counterfactual Fairness over PDAGs

Counterfactual fairness compares the realized outcome for an individual to the counterfactual outcome where a sensitive attribute is changed, but all else stays the same. This notion of fairness is causal-aware and has received considerable research attention for the last few years.

The downside of being causal-aware is that current implementations of counterfactual fairness require knowledge of the causal graph (or better yet, the entire structural model). Here, we demonstrate the existence of bounds for counterfactual fairness with only partial knowledge of the structure, as encoded by a partially-directed acyclic graph (PDAG). We demonstrate these bounds can be found precisely by use of polynomial programming, or estimated efficiently using Bayesian modeling.

- Narrow to “Incentivized counterfactual unfairness”?
    - I think so.

Bounding of Causal Effects despite Structural Uncertainty

Recent work [Zhang, Tian, Bareinboim] outlines how to obtain “Partial Identification” of causal effects, effectively bounding the causal effects even when the effect itself is not identifiable. Here, we apply a similar technique, bounding causal effects where the source of uncertainty regards the structure of the causal model. Specifically, given a partially-directed acyclic graph (PDAG) we provide a general algorithm for bounding counterfactual quantities using canonical representations of SCMs, polynomial programing, and Gibbs sampling. Finally, we explore computational efficiencies obtained by considering only portions of the PDAG, reducing the curse of dimensionality.

Prior over Counterfactuals

As counterfactual reasoning becomes more widespread in machine learning, it becomes increasingly important to identify which priors over counterfactuals are reasonable to use. I postulate that there are “natural” priors on counterfactuals, from which traditional assumptions of minimality in structural learning can follow as a conclusion rather than an assumption. An added benefit is that a prior on counterfactuals could provide a more informative ordering on causal diagrams in an equivalence class, enabling better causal estimates despite structural uncertainty. To distinguish between choices of counterfactual priors, I seek testable implications which could be compared to real-world data.