Seoul National University
Graduate School of Data Science

uai

Beyond Bounds: Quantifying the Probability of Counterfactual Fairness

Taehan Kim, Minyoung Cho, Sanghack Lee

Our Question & Contributions

How likely is a predictor $\widehat{Y}$ to be counterfactually fair?

We estimate the probability that a predictor $\widehat{Y}$ is counterfactually fair under a stated prior, given $P(\mathbf{V})$ and a causal graph $\mathcal{G}$.

- We show that the compatible parameter space is a convex polytope.
- We characterize the counterfactually fair region within the space.
- We estimate the target with Hit-and-Run sampling, complementing prior-free worst-case bounds.

Counterfactual Fairness

Definitions.

- Counterfactual Effect

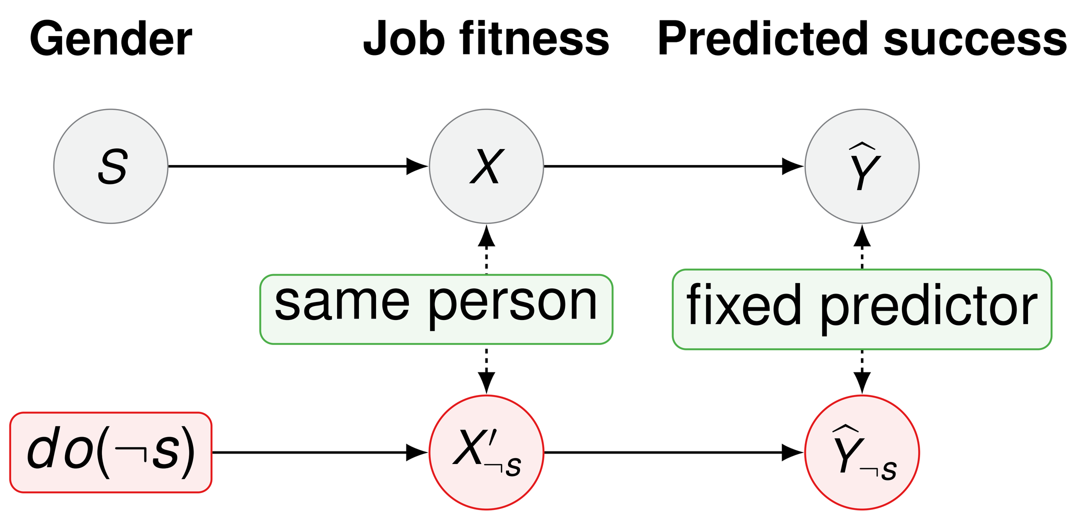
$$\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y}) = P(\hat{y}_{\neg s} \mid \mathbf{x}, s) - P(\hat{y} \mid \mathbf{x}, s).$$

- τ-Counterfactual Fairness

$$\hat{Y} \text{ is } \tau\text{-counterfactually fair} \iff \max_{s, \mathbf{x}, \hat{y}: P(\mathbf{x}, s) > 0} |\mathrm{CE}(s, \neg s \mid \mathbf{x}; \hat{y})| \leq \tau.$$

Example.

- $\widehat{Y}$ is counterfactually fair with respect to gender (S) if its prediction remains unchanged under a counterfactual change in gender.



Fairness Probability over Parameter Space

We adopt canonical SCMs to represent candidate causal mechanisms and map models compatible with $\langle P(\mathbf{V}), \mathcal{G} \rangle$ to a d-dimensional space.

Canonical SCM.[1]

- A Structural Causal Model (SCM) is a 4-tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$:
  - **U**: exogenous, **V**: endogenous, **F**: mechanisms
- A canonical SCM (C-SCM) is defined on finite **V**, where **U** indexes deterministic response functions **F**.
  - C-SCMs can represent all counterfactual distributions.

Parameterizing C-SCMs.

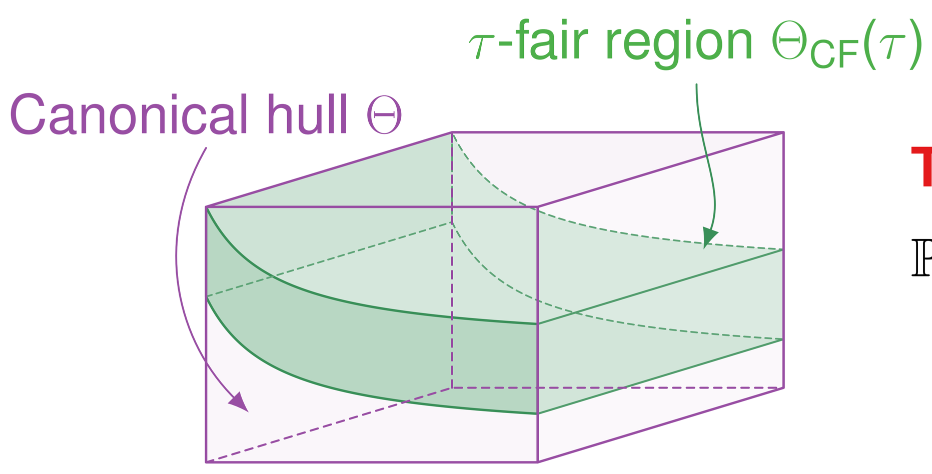
💡 Compatible C-SCMs with the same $\langle P(\mathbf{V}), \mathcal{G} \rangle$ differ only in $P(\mathbf{U})$

- Represent exogenous $P(\mathbf{U})$ by maximal C-component $\{\mathbf{C}_1, \ldots, \mathbf{C}_n\}$:

$$\theta_{\mathcal{M}} = [P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}_1}), \ldots, P_{\mathcal{M}}(\mathbf{u}_{\mathbf{C}_n})]^\top \in \mathbb{R}^d$$

Parameter Space.

- **Canonical hull.** Θ is the set of $\theta_{\mathcal{M}}$ compatible with $\langle P(\mathbf{V}), \mathcal{G} \rangle$.
- **Fair region.** $\Theta_{\mathrm{CF}}(\tau)$ is the set of all $\theta_{\mathcal{M}}$ such that $\widehat{Y}$ is τ-CF.



Target Estimand.

$$\mathbb{P}_{\Theta}(\widehat{Y} \text{ is } \tau\text{-CF}) = \mathbb{P}_{\Theta}(\theta \in \Theta_{\mathrm{CF}}(\tau))$$

Estimating Fairness Probability

Three-step strategy: 1. Characterize the canonical hull. **2. Reduce** to fairness-relevant dimensions. **3. Sample** by Hit-and-Run.

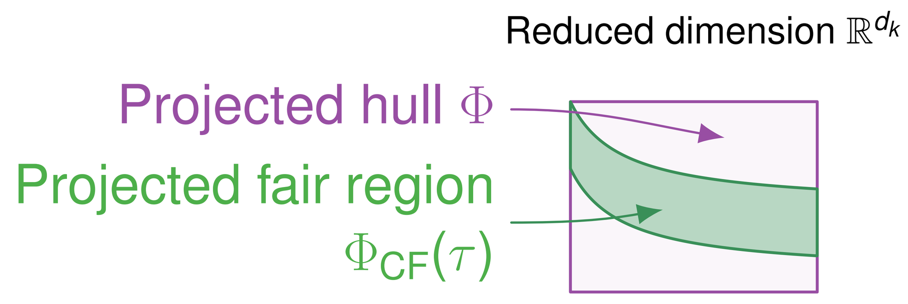
① Characterizing the Canonical Hull

- A canonical hull is a convex polytope. $A_1 \cdot \theta = b_1, \ A_2 \cdot \theta \leq b_2$

② Dimension Reduction

- Evaluate the target in fairness-relevant dimensions.
- Retain only C-components intersecting descendants of S:

$$\phi_{\mathcal{M}} = \pi(\theta_{\mathcal{M}}) \in \mathbb{R}^{d_k}, \quad \pi: \text{ coordinate projection}$$



Reduced-Space Estimand.

$$\mathbb{P}_{\Phi}(\widehat{Y} \text{ is } \tau\text{-CF}) = \mathbb{P}_{\Phi}(\phi \in \Phi_{\mathrm{CF}}(\tau))$$

The same fairness event is evaluated in a lower dimension.
A full-space prior induces $\mathbb{P}_{\Phi} = \pi_{\#}\mathbb{P}_{\Theta}$.

③ Hit-and-Run Sampling

- The target estimand is analytically intractable.
- Since the projected hull Φ remains a convex polytope, we employ the Hit-and-Run algorithm to estimate the probability.
  - Given a declared density on Φ, initialize at $\phi^{(0)} \in \Phi$ and repeat:

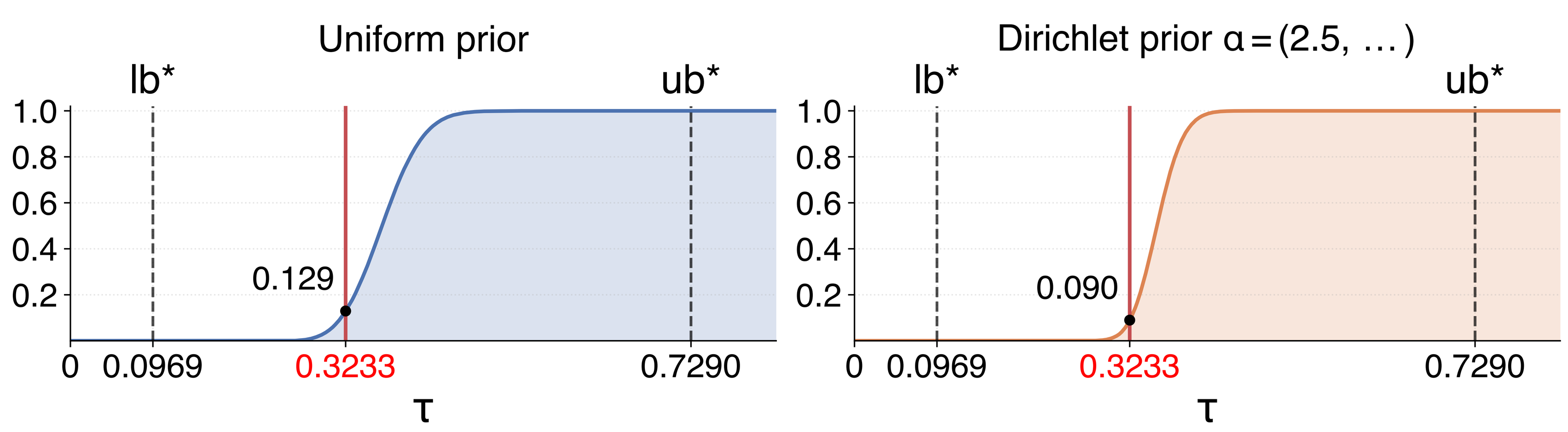
Current point $\phi^{(t)} \in \Phi$ → Sample affine direction d → Compute feasible chord in Φ → Draw next point $\phi^{(t+1)}$

  - For retained draws $\phi^{(1)}, \ldots, \phi^{(T)}$,

$$\widehat{\mathbb{P}}_{\Phi}\left(\widehat{Y} \text{ is } \tau\text{-counterfactually fair}\right) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left\{\phi^{(t)} \in \Phi_{\mathrm{CF}}(\tau)\right\}.$$

Simulation Studies

We evaluate the synthetic dataset $\mathcal{D}_1$ using 10^5 retained samples under uniform and joint Dirichlet ($\alpha = 2.5$) priors.



- For both priors, the transition region $0 < \widehat{\mathbb{P}} < 1$ lies within $[lb^*, ub^*]$ and contains the ground-truth max $|\mathrm{CE}|$ value (red lines).
- The probability curves compare how fair-model mass accumulates under the two priors.
- Our method quantifies the probability that $\widehat{Y}$ is τ-counterfactually fair, providing actionable information for decision-makers.

[1] Zhang, J., Tian, J., and Bareinboim, E. (2022). *Partial Counterfactual Identification from Observational and Experimental Data.* ICML.