Counterfactual Fairness in the Actual World

Alan Mishler (amishler@stat.cmu.edu), Edward H. Kennedy

Dept. of Statistics & Data Science, Carnegie Mellon University

Background

Previous proposal: Construct a fair predictor by performing inference in a fair world that's close in distribution to the actual world.

> *"...just as causal inference is interested in hypothetical worlds representing randomized trials, so is fair inference interested in hypothetical worlds representing fair situations." [1]*

Results:

- Inference in fair world isn't **sufficient** to yield a fair predictor.
- Inference in fair world isn't **necessary** to yield a fair predictor.

Notation, Inference Problem, Research Question

Notation:

Observable variables

$(X,Y) \sim \mathbb{P}$

$A \in X$ = Sensitive attribute like race, gender

$V = X \setminus A$ = Other covariates

Y = Outcome

Potential outcomes

$Y^{A=a} = Y(do(A=a))$

$Y^{aC^{a'}} = Y(do(A=a), C(do(A=a')))$

Assumptions:

- $\mathbb{P}(X,Y)$ generated by known causal DAG.
- Certain (identified) Path-Specific Effect(s) (PSEs) $\psi(\mathbb{P})$ considered unfair a priori, e.g. $A \to Y$.

Proposed fair predictor:

First, define a nearby world $\mathbb{P}^*$ that is (nearly) fair:

$$\mathbb{P}^* = \underset{Q \in \mathcal{Q}}{\arg\min}\, d(\mathbb{Q}, \mathbb{P}) \text{ subject to } \varepsilon_\ell \le \psi(\mathbb{Q}) \le \varepsilon_u \tag{1}$$

for some distance d, e.g. $d(\mathbb{Q},\mathbb{P}) = KL(\mathbb{P}||\mathbb{Q})$ and tolerance $[\varepsilon_\ell, \varepsilon_u]$.

Then the estimand and estimator are:

$$h^*(x) = \underset{h \in \mathcal{H}}{\arg\min}\, \mathbb{E}^*[\ell(h(X), Y)] \tag{2}$$

$$h(x) = \widehat{\underset{h \in \mathcal{H}}{\arg\min}}\, \mathbb{E}^*[\ell(h(X), Y)] \tag{3}$$

for some loss function ℓ, where $\mathbb{E}^*$ is taken in $\mathbb{P}^*$.

Research questions:

Does this procedure yield a **fair** predictor?

Does this procedure yield an **optimal** fair predictor?

Fairness Definitions

Definition type	Name	Definition
Observable	Statistical Parity	$\hat{Y} \perp\!\!\!\perp A$
	Equal opportunity (for classifiers)	$\hat{Y} \perp\!\!\!\perp A \mid Y=1$
	Equal odds	$\hat{Y} \perp\!\!\!\perp A \mid Y$
	Equal accuracy	$\mathbb{P}(\hat{Y}=Y\mid A) = \mathbb{P}(\hat{Y}=Y)$
	False Positive Rate balance	$\mathbb{P}(\hat{Y}=1\mid Y=0,A) = \mathbb{P}(\hat{Y}=1\mid Y=0)$
	False Negative Rate balance	$\mathbb{P}(\hat{Y}=0\mid Y=1,A) = \mathbb{P}(\hat{Y}=0\mid Y=1)$
	Predictive parity	$\mathbb{P}(Y=1\mid \hat{Y}=1,A) = \mathbb{P}(Y=1\mid \hat{Y}=1)$
Causal	Counterfactual fairness *[2]*	$\mathbb{P}(\hat{Y}^a = y\mid V,A) = \mathbb{P}(\hat{Y}^{a'} = y\mid V,A)$
	No unresolved discrimination *[3]*	No path $A \to \ldots \to Y$ not mediated by a resolving variable.
	No proxy discrimination *[3]*	No path $A \to \ldots \to Y$ through a prohibited proxy variable.
	Path-specific fairness on $\hat{Y}$	No prohibited PSEs terminating in $\hat{Y}$

etc.

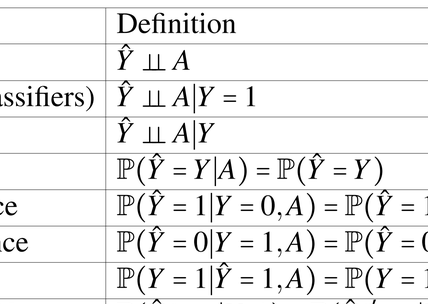
Inference in fair world $\Longrightarrow$ fair predictor

Example 1: The real world is fair, so $\mathbb{P} = \mathbb{P}^*$ *[2]*

A = race or gender
$C = \mathbb{1}\{\text{drives a red car}\}$
$U = \mathbb{1}\{\text{likes aggressive driving}\}$ (unobserved)
Y = accident rate

The real world is fair (no paths $A \to \ldots \to Y$); but an unconstrained predictor **is not**.

Let $\hat{Y} = h(C,A) = E[Y\mid C,A]$, the MSE-optimal predictor.



Inference in fair world $\Longrightarrow\!\!\!\!\!\!\not\quad$ fair predictor cont.

Example 1 continued

Total effect of A on $\hat{Y}$:

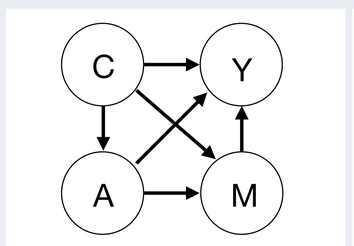
$$\begin{aligned}\mathbb{E}[h(C^1,1) - h(C^0,0)] &= \mathbb{E}\{\mathbb{E}[h(C,1)\mid A=1] - \mathbb{E}[h(C,0)\mid A=0]\}\\ &= \mathbb{E}\{\mathbb{E}[Y\mid C,A=1]\mid A=1\} - \mathbb{E}\{\mathbb{E}[Y\mid C,A=0]\mid A=0\}\\ &= \mathbb{E}[Y\mid A=1] - \mathbb{E}[Y\mid A=0]\\ &= 0\end{aligned}$$

Natural direct effect of A on $\hat{Y}$:

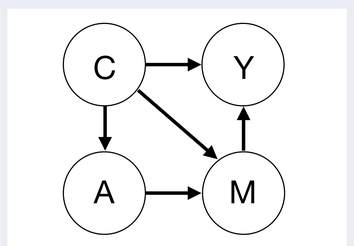
$$\begin{aligned}\mathbb{E}[h(C^1,1) - h(C^1,0)] &= \mathbb{E}\{\mathbb{E}[h(C,1) - h(C,1)\mid A=1]\}\\ &= \mathbb{E}\{\mathbb{E}[Y\mid C,A=1] - \mathbb{E}[Y\mid C,A=0]\mid A=1]\}\\ &= \mathbb{E}[Y\mid A=1] - \mathbb{E}\{\mathbb{E}[Y\mid C,A=0]\mid A=1]\}\\ &\neq 0 \text{ in general}\end{aligned}$$

Result: $\hat{Y}$ isn't fair if we care about NDE, or by observational fairness definitions.

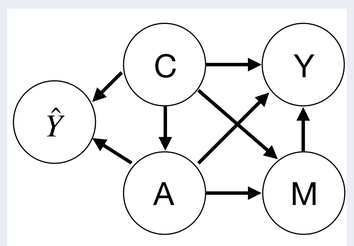
Example 2: The real world is unfair, so $\mathbb{P} \neq \mathbb{P}^*$



(a) Causal DAG of real world $\mathbb{P}$.



(b) Causal DAG of fair world $\mathbb{P}^*$.



(c) Causal DAG with predictor $\hat{Y} = h(C,A)$

Suppose the natural direct effect of A on Y is prohibited:

$$\begin{aligned}\mathbb{E}^*[Y^{1M^1} - Y^{0M^1}] &= \mathbb{E}^*_C\{\mathbb{E}^*_M[\mathbb{E}^*[Y\mid C,M,A=1] - \mathbb{E}^*[Y\mid C,M,A=0]\mid C,A=1]\}\\ &= 0 \text{ by construction}\end{aligned}$$

Let $\hat{Y} = h(C,A) = \mathbb{E}[Y\mid C,A]$, the MSE-optimal predictor. *[1]*

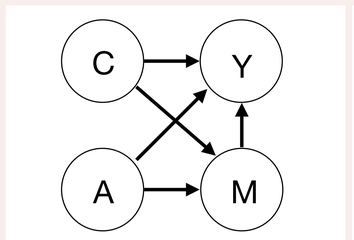
Natural direct effect of A on $\hat{Y}$:

$$\begin{aligned}\mathbb{E}[h(C,1) - h(C,0)] &= \mathbb{E}_C\{\mathbb{E}^*[Y\mid C,A=1] - \mathbb{E}^*[Y\mid C,A=0]\}\\ &\neq 0 \text{ in general}\end{aligned}$$

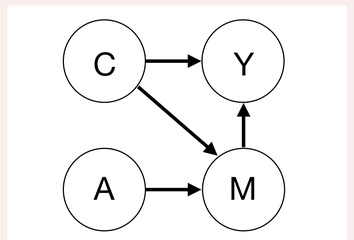
Result: $\hat{Y}$ isn't guaranteed to be fair by any of the above definitions.

Inference in fair world $\Longleftarrow\!\!\!\!\!\!\not\quad$ fair predictor

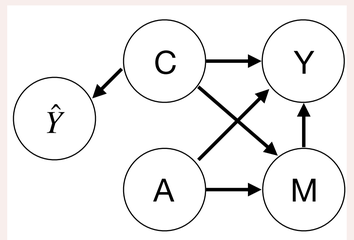
Example 3



(a) Causal DAG of real world $\mathbb{P}$.



(b) Causal DAG of fair world $\mathbb{P}^*$.



(c) Causal DAG with predictor $\hat{Y} = h(C)$

$\hat{Y}$ trivially satisfies causal fairness definitions (though not observational definitions).

E.g. for MSE loss ℓ, optimal predictor on C is $\hat{Y}_{\text{opt}} = \mathbb{E}[Y\mid C]$.
Vs. a predictor derived in the fair world: $\hat{Y}^* = \mathbb{E}^*[Y\mid C]$.

Both predictors are trivially fair, but the loss relationship is:

$$\mathbb{E}[\ell(\hat{Y}_{\text{opt}}, Y)] = \mathbb{E}[\ell(\mathbb{E}[Y\mid C], Y)] \le \mathbb{E}[\ell(\mathbb{E}^*[Y\mid C], Y)] = \mathbb{E}[\ell(\hat{Y}^*, Y)] \tag{4}$$

The middle inequality is almost certainly strict, since $\mathbb{E}^*[Y\mid C] \neq \mathbb{E}[Y\mid C]$.

Result: Restricting $h(C)$ to a functional on $\mathbb{P}^*$ unnecessarily **increases risk.**

Path-specific fairness and decision making

Fairness in terms of PSEs ending in Y doesn't make sense unless $\hat{Y}$ has an effect on Y.

Option 1: Consider fairness in terms of PSEs terminating in $\hat{Y}$.

$$\hat{Y} = h(x) = \underset{h \in \mathcal{H}}{\arg\min}\, \mathbb{E}^*[\ell(h(X), Y)] \text{ subject to } \varepsilon_\ell \le PSE(A \text{ to } H) \le \varepsilon_u \tag{5}$$

Option 2: Consider the effect of $\hat{Y}$ on Y.

Proposal: Fairness is a relative measure of the extent to which a predictor in context makes the world fairer (future work)

References

[1] Razieh Nabi & Ilya Shpitser (2018). Fair Inference on Outcomes. http://arxiv.org/abs/1705.10378.

[2] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. 2017. http://arxiv.org/abs/1703.06856.

[3] Niki Kilbertus, et al. Avoiding Discrimination through Causal Reasoning. Advances in Neural Information Processing Systems, (30): 1-11, 2017.