Dressing Room

David Reber
Columbia University
david.reber@columbia.edu

Abstract: abstract

1. Tentative outline of Mid-semester report

TODO: create outline of what I want in the mid-semester report
TODO: remove spurious notes from this doc
Literature review:

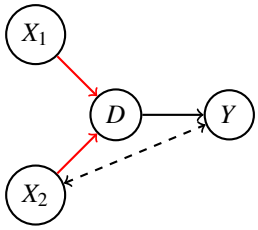
- causal fairness
- Causal Explanation Forumula
- (Is there a paper for the heirarchy we saw in lecture?)
- ...
- Agent Incentives: A Causal Perspective
- incentivized Unfairness: how fair labels can yield unfair predictions
-

Confident Content:

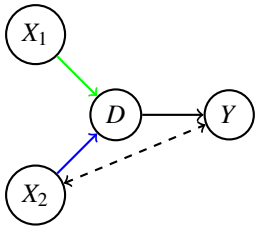
- Notation
- Definitions
- a class of graphs which are non-ID in G but ID in G^{min}, with proof
- each of the 'unincentivzed null' theorems, with proofs (articulating this will probably help me articulate what I mean by incentivized effects)
-

Speculative Content:

- proving that incentives decompose the $Cft\text{-}DE$ effect
- my conjecture for evidencing rewards?
- expressing $Cft\text{-}DE^{\uparrow}$ as a v'-specific effect
- articulating what unit-level incentives look like (the equivalent of the unit-specific measures in the heirarchy). For that matter, is there a pure structural version of incentivized unfairness? (just expressed in terms of the deterministic counterfactuals, with no probability distribution?)
- proving the Fair Prediction Theorem for introduced total variation (using incentivized Cft-DE, etc).
-



(a) Direct effect of $\mathbf{X} = \{X_1, X_2\}$ on D



(b) Decomposition of the direct effect of $\mathbf{X}$ on D

Figure 1: A simple example of an incentives decomposition of the direct effect of $\mathbf{X}$ on D, as rewarded by Y.

2. First steps on non-markovian response incentives

The following notation and definitions help keep consistency with the lab's existing work on both fairness [**?**] and CRL [**?**] (see Figure 1a for a visual aid). In addition to the usual SCM $M = (V, U, F, P(U))$, we use Y as the reward variable(s), $D \in An(Y)$ as the decision node, and $X \in An(D)$ as the sensitive attribute(s) which the decision may be unfair towards. $\mathbf{C} \subset V$ are the covariates which D is allowed to observe; that is, $Pa(D) \subset \mathbf{C}$. A policy π is a soft intervention on D which respects $domain(\pi) \subset \mathbf{C} \cup \{U_D\}$, and an *optimal policy* is defined as a policy π that maximizes the sum of the expected rewards: $\mathbb{E}_\pi[\sum_{Y \in \mathbf{Y}} Y]$. Lastly, a 'fairness effect E' is just some element $E \in \{Cft\text{-}DE,\ Cft\text{-}IE,\ Cft\text{-}SE\}$.

TODO: update to more general causal fairness notions

Definition 1 (Response Incentive) *Let $M = (V, U, F, P(U))$ be an SCM. A policy π responds to a variable $X \in V$ if there exists some intervention $do(X = x)$ and some exogenous unit $U = u$, such that $D_x(u) \neq D(u)$. The variable X has an* response incentive *if all optimal policies respond to X. We say a causal diagram G* admits *an response incentive on X if it is compatible with an SCM that has an response incentive on X.*

Note that *response incentive* is a binary classification of the ancestors of D: either a node has an response incentive, or it doesn't. Next, the *minimal reduction* of a causal graph G eliminates all edges from parents of D which do not have their own d-connection to Y.

Definition 2 (Minimal Reduction) *The* minimal reduction G^{min} *of a causal diagram G is the result of removing from G all edges $X \to D$ satisfying $(X \perp\!\!\!\perp Y | D \cup Pa(D) \setminus X)$.*

Intuitively, the minimal reduction breaks all non-informative links to the decision node. The main result from [1] we build on is a graphical criterion for classifying a node X as having an response incentive.

Conjecture 1 (Response Incentive Criterion [1]) *Let G be an acyclic causal diagram. Then G admits a response incentive on $X \in V$ if and only if the minimal reduction G^{min} has a directed path X - $\to D$.*

2.1. *Gathering thoughts, Next steps*

- the soundness portion is proved in the main text, and relies on Lemma 25 (which in turn relies on Lemmas 24, 21,).
- the completeness direction is Lemma 28, which is lengthy but does not depend on any other lemmata.
- the so-called completeness direction says 'If the graphical criteria holds, then there is a response incentive on X in at least one SCM compatible with G'.
- ...I don't think the completeness direction is the one I care as much about: it says 'If I say there's a response incentive then there probably is', whereas the soundess direction says 'if I don't say there's a response incentive, then there definately isn't'
- So I think the soundness direction is the one I want more: 'If the graphical condition does not hold for X, then X does not have a response incentive for any SCM'.

First step: Just translate the soundness direction from the text (recorded below) precisely with my notation.

Second step: Pretend the supporting Lemmata (21,24,25) all generalize to the Markovian case, and attempt to generalize this portion from the text.

Backup second: if the second step fails, try just generalizing Lemmas 21, 24, and 25, in that order.

2.2. *Original Proof (Markov case)*

The *if* (completeness) direction is proved in Lemma 28 in Appendix C.2. For the soundness direction, assume that for G, the minimal reduction G^{min} does not contain a directed path $X \dashrightarrow D$. Let $M = (G, E, \mathbf{F}, \mathbf{P})$ be any SCIM compatible with G. Let $M^{min} = (G^{min}, E, \mathbf{F}, \mathbf{P})$ be M, but with minimal reduction G^{min}. By Lemma 25 in Appendix C, there exists a G^{min}-respecting policy $\tilde{\pi}$ that is optimal in M. In $M^{min}_{\tilde{\pi}}$, X is causally irrelevant for D so $D_x(u) = D(u)$. Furthermore, $M_{\tilde{\pi}}$ and $M^{min}_{\tilde{\pi}}$ are the same SCM, with the functions $\mathbf{F} \cup \tilde{\pi}$. So $D_x(u) = D(u)$ also in $M_{\tilde{\pi}}$, which means that there is an optimal policy in M that does not respond to interventions on X for any ε.

2.3. *In-text portion of soundness proof, translated to my notation*

Theorem 1 (Response Incentive Criterion [1]) *Let G be a Markovian causal diagram. Assume the minimal reduction G^{min} does not have a directed path $X \dashrightarrow D$. Then G does not admit a response incentive on $X \in V$.*

Proof: Assume that for G, the minimal reduction G^{min} does not contain a directed path $X \dashrightarrow D$ (perhaps needs to be stronger for non-Markovian?). Let M be any SCM compatible with G. Partition $\mathbf{Pa}_D$ into the non-requisite parents $\mathbf{Pa}^{non}_D = \{W \in \mathbf{Pa}_D : (W \perp\!\!\!\perp Y | D \cup Pa(D) \setminus W)\}$ and requisite parents $\mathbf{Pa}^{req}_D = \mathbf{Pa}_D \setminus \mathbf{Pa}^{non}_D$. (Note that by definition of G^{min}, $\mathbf{Pa}^{req}_D$ is precisely the set of D's parents in G^{min}). By Lemma 25 (requires Markovian?), there exists a G^{min}-respecting policy $\tilde{\pi}$ that is optimal in M. Let $M_{\tilde{\pi}} := M_{\sigma_D}$ where $\sigma_D = \tilde{\pi}(D|\mathbf{Pa}^{req}_D)$. Since X is causally irrelevant in $M_{\tilde{\pi}}$, we have $D_x(u) = D(u)$. This means that there is an optimal policy in M that does not respond to interventions on X for any U. ...which I think means there's no response incentive on X?

Lemma 21 is just Rule 1 of the do-calculus, which holds for Non-Markovian.

Lemma 24 is the intersection property of d-seperation, which holds for Non-Markovian.

So only Lemma 25 and the text body need to be checked.

2.4. *Ancestry is Necessary for Response Incentive*

The following theorem is probably proved somewhere already (TODO: look) but it was a good exercise.

Theorem 2 (Invariance of non-descendents to interventions) *Assume $X \cap An(D) = \emptyset$. Then $D_x(u) = D(u)$ for all $x \in X$, $u \in U$.*

We use the following lemma:

Lemma 3 (Invariance Inheritance) *Assume $X \cap An(D) = \emptyset$ and $Pa(D)_x(u) = Pa(D)(u)$. Then $D_x(u) = D(u)$.*

Proof of Lemma 3: Let $u \in U$, $x \in X$. For convienence of notation, let $Q := Pa(D)(u)$, the value that $Pa(D)$ realizes when $U = u$.

Then

$$\begin{aligned} D(u) &= D_Q(u) && \text{Consistency} \\ &= D_{Q,x}(u) && \text{Exclusion Restrictions, since } X \cap Pa(D) = \emptyset \\ &= D_x(u) && \text{Exclusion Restrictions, since } Pa(D)_x(u) = Pa(D)(u) \end{aligned}$$

Proof of Theorem 2: Let N^i be the set of D's i^{th} grandparents; that is, $N^0 = Pa(D)$, and $N^{i+1} = Pa(N^i) \setminus N^i$. Let l be the largest index corresponding to a non-empty N^i, that is $N^l \neq \emptyset$ and $N^{l+1} = \emptyset$. (So long as $D \neq \emptyset$, l exists and is unique). Note that $\{N^0, ..., N^l\}$ forms a partition of $An(D)$.

We apply Lemma 3 inductively in reverse order: *Base case:* Since for all $W \in N^l$ we have $Pa(W) = \emptyset$, we have $W(u) = W_x(u)$ trivially by Exclusion Restrictions. Thus $N^l(u) = N^l_x(u)$. *Inductive step:* Assume $N^{l-k}(u) = N^{l-k}_x(u)$. Let $W \in N^{l-k-1}$. Since $Pa(W) \subset N^{l-k}$, we have that $Pa(W)_x(u) = Pa(W)(u)$, so by Lemma 3, $W_x(u) = W(u)$. Thus $N^{l-k-1}(u) = N^{l-k-1}_x$.

By applying the inductive step $l-1$ times after the base case, we obtain that $D_x(u) = D(u)$ as desired.

Corollary 1 (Necessity of Ancestry) *Let G be an acyclic causal diagram which admits a response incentive on some variable $X \in V$. Then $X \in An(D)$.*

The contrapositive of the corollary follows directly from Theorem 2, since all policies (including optimal ones) are invariant to interventions on X if $X \notin An(D)$.

2.5. *Non-Markovian Soundness*

Theorem 4 (Non-Markovian Response Incentive Criterion) *Let G be an acyclic causal diagram. Assume the minimal reduction G^{min} does not have a directed path $X \dashrightarrow D$. Then G does not admit a response incentive on $X \in V$.*

Proof: Assume that for G, the minimal reduction G^{min} does not contain a directed path $X \dashrightarrow D$. Let M be any SCM compatible with G. Partition $\mathbf{Pa}_D$ into the non-requisite parents $\mathbf{Pa}_D^{non} = \{W \in \mathbf{Pa}_D : (W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)\}$ and requisite parents $\mathbf{Pa}_D^{req} = \mathbf{Pa}_D \setminus \mathbf{Pa}_D^{non}$. (Note that by definition of G^{min}, $\mathbf{Pa}_D^{req}$ is precisely the set of D's parents in G^{min}). By Lemma 25 (requires Markovian?), there exists a G^{min}-respecting policy $\tilde{\pi}$ that is optimal in M. Let $M_{\tilde{\pi}} := M_{\sigma_D}$ with $\sigma_D = \tilde{\pi}(D|\mathbf{Pa}_D^{req})$. Since X is causally irrelevant in $M_{\tilde{\pi}}$, we have $D_x(u) = D(u)$. This means that there exists an optimal policy intervention on M that does not respond to interventions on X for any U, so X does not have a response incentive relative to the SCM M. Since M was arbitrary, G does not admit a response incentive on X.

(All I did was translate to Elias language, no changes were required for non-markovian).

2.6. *Lemma 25 work*

Lemma 5 (Lemma 25 copied: Gmin-respecting optimal policy) *Every single-decision SCIM $M = (G, E, F, P)$ has an optimal policy $\tilde{\pi}$ that depends only on requisite observations. In other words, $\tilde{\pi}$ is also a policy for the minimal model $M^{min} = (G^{min}, E, F, P)$. We call $\tilde{\pi}$ a G^{min}-respecting optimal policy.*

Proof: TODO (when translate): move up defn of $\mathbf{Y}_D$. Standardize U meaning utility, or find different notation.

First partition $\mathbf{Pa}_D^G$ into the non-requisite parents $\mathbf{Pa}_D^{non} = \{W \in \mathbf{Pa}_D : (W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)\}$ and requisite parents $\mathbf{Pa}_D^{req} = \mathbf{Pa}_D^G \setminus \mathbf{Pa}_D^{non}$.

Let π^* be an optimal policy in M. To construct a G^{min}-respecting version $\tilde{\pi}$, select any value $\tilde{\mathbf{pa}}_D^{non} \in dom(\mathbf{Pa}_D^{non})$ for which $Pr_{\pi^*}(\mathbf{Pa}_D^{non} = \tilde{\mathbf{pa}}_D^{non}) > 0$. For all $\mathbf{pa}_D^{req} \in dom(\mathbf{Pa}_D^{req})$ and $\varepsilon_D \in dom(E^D)$, let

$$\tilde{\pi}(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}, \varepsilon_D) := \pi^*(\mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}, \varepsilon_D)$$

The policy $\tilde{\pi}$ is permitted in M^{min} because it does not vary with $\mathbf{Pa}_D^{non}$.

Now let us prove that $\tilde{\pi}$ is optimal in M. Partition $\mathbf{Y}$ into $\mathbf{Y}_D = \mathbf{Y} \cap \mathbf{Desc}_D$ and $\mathbf{Y}_{\setminus D} = \mathbf{Y} \setminus \mathbf{Desc}_D$. D is causally irrelevant for every $Y \in \mathbf{Y}_{\setminus D}$ so every policy π (in particular, $\tilde{\pi}$) is optimal with respect to $U^{\setminus D} = \sum_{Y \in \mathbf{Y}_{\setminus D}} Y$.

We now consider $\mathbf{Y}_D$. By definition, $(W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)$ for every $W \in \mathbf{Pa}_D^{non}$. By inductively applying the intersection property of d-separation (Lemma 24) over elements of $\mathbf{Pa}_D^{non}$ we obtain

$$\mathbf{Pa}_D^{non} \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D^{req} \tag{1}$$

Next we establish that $\mathbb{E}_{\tilde{\pi}}[U_D] = \mathbb{E}_{\pi^*}[U_D]$ by showing that $\mathbb{E}_{\tilde{\pi}}[U_D|\mathbf{pa}_D] = \mathbb{E}_{\pi^*}[U_D|\mathbf{pa}_D]$ for every $\mathbf{pa}_D \in dom(\mathbf{Pa}_D)$ with $\Pr(\mathbf{pa}_D) > 0$. First, the expected utility of $\tilde{\pi}$ given any $(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non})$ with $\Pr(\mathbf{Pa}_D^{req} = \mathbf{pa}_D^{req}, \mathbf{Pa}_D^{non} = \mathbf{pa}_D^{non}) > 0$ is equal to the expected utility of π^* on input $(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non})$:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}}[U_D|\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}] \\
&\qquad = \sum_{u,d} \left(u \Pr(U_D = u | d, \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}) \cdot \Pr_{\tilde{\pi}}(D = d | \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}) \right) \\
&\qquad = \sum_{u,d} \left(u \Pr(U_D = u | d, \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}) \cdot \Pr_{\pi^*}(D = d | \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}) \right) \\
&\qquad = \mathbb{E}_{\pi^*}[U_D|\mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}]
\end{aligned}$$

where the middle equality follows from (1) and the definition of $\tilde{\pi}$. Second, the expected utility of π^* given input $\tilde{\mathbf{pa}}_D^{non}$ is the same as its expected utility on any input $\mathbf{pa}_D^{non}$:

$$\begin{aligned}
&= \max_d \mathbb{E}_{\pi^*}[U_D^d|\mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}] \\
&= \max_d \mathbb{E}_{\pi^*}[U_D^d|\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}] \\
&= \mathbb{E}_{\pi^*}[U_D|\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}]
\end{aligned}$$

where the first equality follows from the optimality of π^* and the second from Lemma 21. The expression $\mathbb{E}_{\pi^*}[U_D^d|\ldots]$ means that we first assign the policy π^* then intervene to set $D = d$, which renders π^* effectively irrelevant but formally necessary for creating an SCM. This result shows that $\tilde{\pi}$ is optimal for U_D and has $\mathbb{E}_{\tilde{\pi}}[U_D] = \mathbb{E}_{\pi^*}[U_D]$. Since $\tilde{\pi}$ is optimal for both U_D and $U_{\setminus D}$, $\tilde{\pi}$ is optimal in M.

Lemma 6 (Lemma 25 translated: Gmin-respecting optimal policy) *For every single-decision, acyclic SCM $M = (V, U, F, P(U))$ there exists an optimal policy intervention $\tilde{\pi}$ on D that depends only on requisite observations. In other words, $M_{\tilde{\pi}}$ is compatible with G^{min}. We call $\tilde{\pi}$ a G^{min}-respecting optimal policy.*

Proof: First partition $\mathbf{Y}$ into $\mathbf{Y}_D = \mathbf{Y} \cap \mathbf{Desc}_D$ and $\mathbf{Y}_{\setminus D} = \mathbf{Y} \setminus \mathbf{Desc}_D$. Also partition $\mathbf{Pa}_D^G$ into the non-requisite parents $\mathbf{Pa}_D^{non} = \{W \in \mathbf{Pa}_D : (W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)\}$ and requisite parents $\mathbf{Pa}_D^{req} = \mathbf{Pa}_D^G \setminus \mathbf{Pa}_D^{non}$.

Let π^* be an optimal policy in M. To construct a G^{min}-respecting version $\tilde{\pi}$, select any value $\tilde{\mathbf{pa}}_D^{non} \in dom(\mathbf{Pa}_D^{non})$ for which $Pr_{\pi^*}(\mathbf{Pa}_D^{non} = \tilde{\mathbf{pa}}_D^{non}) > 0$. For all $\mathbf{pa}_D^{req} \in dom(\mathbf{Pa}_D^{req})$ and $u_D \in dom(U_D)$, let

$$\tilde{\pi}(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}, u_D) := \pi^*(\mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}, u_D)$$

The policy $\tilde{\pi}$ is permitted in M^{min} because it does not vary with $\mathbf{Pa}_D^{non}$.

Now let us prove that $\tilde{\pi}$ is optimal in M. Note that D is causally irrelevant for every $Y \in \mathbf{Y}_{\setminus D}$ so every policy π (in particular, $\tilde{\pi}$) is optimal with respect to $R_{\setminus D} = \sum_{Y \in \mathbf{Y}_{\setminus D}} Y$.

We now consider $\mathbf{Y}_D$. By definition, $(W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)$ for every $W \in \mathbf{Pa}_D^{non}$. By inductively applying the intersection property of d-separation (Lemma 24) over elements of $\mathbf{Pa}_D^{non}$ we obtain

$$\mathbf{Pa}_D^{non} \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D^{req} \tag{2}$$

Next we establish that $\mathbb{E}_{\tilde{\pi}}[R_D] = \mathbb{E}_{\pi^*}[R_D]$ by showing that $\mathbb{E}_{\tilde{\pi}}[R_D | \mathbf{pa}_D] = \mathbb{E}_{\pi^*}[R_D | \mathbf{pa}_D]$ for every $\mathbf{pa}_D \in dom(\mathbf{Pa}_D)$ with $\Pr(\mathbf{pa}_D) > 0$. First, the expected reward of $\tilde{\pi}$ given any $(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non})$ with $\Pr(\mathbf{Pa}_D^{req} = \mathbf{pa}_D^{req}, \mathbf{Pa}_D^{non} = \mathbf{pa}_D^{non}) > 0$ is equal to the expected reward of π^* on input $(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non})$:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}}[R_D | \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}] \\
&\quad = \sum_{u,d} \left(u \Pr(U_D = u | d, \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}) \cdot \Pr_{\tilde{\pi}}(D = d | \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}) \right) \\
&\quad = \sum_{u,d} \left(u \Pr(U_D = u | d, \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}) \cdot \Pr_{\pi^*}(D = d | \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}) \right) \\
&\quad = \mathbb{E}_{\pi^*}[R_D | \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}]
\end{aligned}$$

where the middle equality follows from (2) and the definition of $\tilde{\pi}$. Second, the expected reward of π^* given input $\tilde{\mathbf{pa}}_D^{non}$ is the same as its expected reward on any input $\mathbf{pa}_D^{non}$:

$$\begin{aligned}
&= \max_d \mathbb{E}_{\mathrm{do}(D=d)}[R_D | \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}] \\
&= \max_d \mathbb{E}_{\mathrm{do}(D=d)}[R_D | \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}] \\
&= \mathbb{E}_{\pi^*}[R_D | \mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}]
\end{aligned}$$

where the first equality follows from the optimality of π^* and the second from Rule 1 of the do-calculus. This result shows that $\tilde{\pi}$ is optimal for R_D and has $\mathbb{E}_{\tilde{\pi}}[R_D] = \mathbb{E}_{\pi^*}[R_D]$. Since $\tilde{\pi}$ is optimal for both R_D and $R_{\setminus D}$, $\tilde{\pi}$ is optimal in M.

2.7. Non-Markovian Conclusions

3/1: I've fully convinced myself that the soundness proof already holds for Non-markovian settings (including Lemmas 21, 24, 25). The only changes I made were notational (like staying in SCM land, and never referencing SCIMs). The intuition for why the result already holds in Non-markov setting, is that non-markovianity only affects what G^{min} looks like. But once we have G^{min} (as is assumed by the criteria), the only independence relation we need is guarenteed by definition of G^{min}.

New 3/2 thoughts: I think it may be even simpler than I thought. WLOG we can consider all $\mathbf{U}$ as endogenous variables, because we never use $P(U)$ to compute anything; all we care about are the connections between things. That's why nothing changes once it's all pulled in to G^{min}. Specifically, we allow $\mathbf{Pa}_D$ to include elements of $\mathbf{U}$. (Unless we can guarentee that π^* doesn't vary with $\mathbf{Pa}_D^{non}$, this might pose a problem: in this case we can't fix the value of $\tilde{\mathbf{pa}}_D^{non}$. Not an issue if my conjecture about 'all optimal policies ignore X' is true, but is it?)

Are my 'ancestry is necessary' and the soundness theorem identical, without realizing it? ...Not quite. The soundness condition is a stronger version; the difference is exactly that of counterfactual unfairness and incentivized unfairness. The ancestry condition says 'If there's a response incentive on X then $X \in An(D)$ in G'; the soundness condition says 'If there's a response incentive on X then $X \in An(D)$ in G^{min}'.

Back to the non-markovianity: I think for sure the proof holds, because we don't have to know what $\tilde{\pi}$ is to know that it exists (with all the consequent implications). So the theorem holds in the non-markov case. But what I'm unsure of, and want to come back to, is whether $\tilde{\pi}$ can always be found/approximated in the non-markov case. But this may be outside the scope of my current research.

3. If there's no response incentive on X, every π^* will ignore X.

Conjecture 2 (All optimal policies are fair w.r.t. unincentivized variables) *Let G be an acyclic causal diagram, such that the minimal reduction G^{min} does not have a directed path $X \dashrightarrow D$.*

Let M be an arbitrary SCM compatible with G, and π^* some policy intervention on D which is optimal in M. Then π^* does not respond to X; that is, $\pi^*_{X=x}(u) = \pi^*(u)$ for all $x \in X$, $u \in U$.

The only thing that will be different here (I think), is that instead of simply showing that there exists some optimal policy that has this property, we're going to show that it holds for all optimal policies.

Proof: Assume that for G, the minimal reduction G^{min} does not contain a directed path $X \dashrightarrow D$. Let M be any SCM compatible with G. Let π^* be some policy intervention on D which is optimal in M.

First partition $\mathbf{Pa}_D^G$ into the non-requisite parents $\mathbf{Pa}_D^{non} = \{W \in \mathbf{Pa}_D : (W \perp\!\!\!\perp \mathbf{Y}_D | D \cup \mathbf{Pa}_D \setminus W)\}$ and requisite parents $\mathbf{Pa}_D^{req} = \mathbf{Pa}_D^G \setminus \mathbf{Pa}_D^{non}$.

Following Lemma 25, construct a G^{min}-respecting version $\tilde{\pi}$ by selecting any value $\tilde{\mathbf{pa}}_D^{non} \in dom(\mathbf{Pa}_D^{non})$ for which $Pr_{\pi^*}(\mathbf{Pa}_D^{non} = \tilde{\mathbf{pa}}_D^{non}) > 0$. For all $\mathbf{pa}_D^{req} \in dom(\mathbf{Pa}_D^{req})$ and $u_D \in dom(U^D)$, let

$$\tilde{\pi}(\mathbf{pa}_D^{req}, \mathbf{pa}_D^{non}, u_D) := \pi^*(\mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non}, u_D)$$

By Lemma 25, $\tilde{\pi}$ is optimal in M, and $M_{\tilde{\pi}}$ is compatible with G^{min}, so as previously argued $\tilde{\pi}_{X=x}(u) = \tilde{\pi}(u)$ in M for all $x \in X$, $u \in U$.

Hence, it suffices to show that $\pi^*_{X=x}(u) = \tilde{\pi}_{X=x}(u)$ and $\tilde{\pi}(u) = \pi^*(u)$ in M for all $x \in X$, $u \in U$.

3.1. thoughts on conjecture

I currently (3/1) expect the conjecture to be false. I think changing the choice of $\tilde{\mathbf{pa}}_D^{non}$ might affect the distribution $\Pr_{\tilde{\pi}}(D | \mathbf{pa}_D^{req}, \tilde{\mathbf{pa}}_D^{non})$, but in such a way that cancels out when you take the expecation? **I think the best way forward here is to look for a counter-example that does this.** Current guess of where to start looking for such a counterexample is with confounding $W \rightarrow X$, where W is a mediator (simply because I know this situation is non-ID for Cft-measures).

4. Applications to Cft-decomposition

I conjecture confidently that a purely-spurious effect will never be incentivized. I'm not sure how a decomposition of incentives would work...Perhaps the paper Ryan sent me will be helpful.

It should be straightforward to find examples where direct- and indirect-effects are ID in G^{min} but not in G: just make sure the non-ID subgraph is d-seperated from Y given D, $Pa(D)$.

After you have some ID examples, construct a step-by-step of the 'algorithm' for determining that $Cft - DE, IE^{\uparrow}$ are ID (Primarily to help others like my mentors pick it up quicker).

5. Reading Introduced Unfairness

Questions:

So is it correct to say that feeding a sensitive feature to the algorithm (while it does guarentee it won't introduce new unfairness) it prevents the algorithm from reducing that unfairness? Yes, this is true: see Theorem 13 .

So I think $\hat{Y}$ is my D? And U is my Y? And we're trying to maximize U.

I don't see how inputs to $\hat{Y}$ can be descendants of Y.

"Absence of separation means that the model has added a dependence between A and Yˆ , that was not present between A and Y." I don't get it. In Figure 1, this doesn't seem to hold.

I think theorem 11 is relevant for me: all those situations I've been drawing where X is only connected to D via a bidirected edge to a requisite parent, for instance: in these cases, there is no response incentive on X, but there is an incentive for introduced total variation. How are these compatible?? I'll bet it's through the causal explanation formula: that the increase in total variation is only through the spurious effect, while still satisfing $D_x(u) = D(u)$.

A P-admissible loss function seems to be defined only relative to a given SCM; my guess is that some loss functions are always P-admissible, no matter the SCM (such MSE)?

"In cases where the labels are independent of A, the only useful information A might provide is how the predictor should interpret its features. Since adding A as an observation prevents ITV, we can infer that part of the information that A provides is how to interpret its influence on the available features. That is, knowing A would help the predictor disentangle information about A and Y . In summary, predictors with P-admissible loss can become unfair in spite of fair labels because they are unable to disentangle information about A and Y". Damn, that's hilarious! ...but also, maybe it makes sense - I can imagine a human getting frustrated because it's trying to be fair, but all they're being fed are spurious variables known to be correlated with the sensitive attribute, and they can't know how to be fair as a result.

Good summary sentence: “This challenges the notion of “fairness through unawareness”, as it suggests that making the sensitive attribute available as a feature can improve fairness **when labels are fair**.”

“...since ITV = 0 is a specific group-level measure, it does not come with individual-level guarantees. In the music test example, as the initial test has lower accuracy for women, women who pass the initial test receive a slightly lower prediction when A is used explicitly compared to when it is not (0.903 instead of 0.905, see Table 1). Even though this negative effect is offset by the higher score given to women who failed the test (0.14 instead of 0.1), this may still be perceived as unfair by the high aptitude women who passed the test T.”

“For instance, if mean squared error is used to produce $\hat{Y} = p \in [0, 1]$, but a binary accept/reject is required, then thresholding (e.g. at 0.5) reduces to the zero-one loss case, and may give ITV > 0, even if the Theorem 13 criteria are met. In this case, randomising the result (accepting with probability p) preserves the result. However, our results do not rely on randomness in general.”

I like the idea of mimicing the dataset-usage of this paper: simulations on random graphs generated by PyCID, and the Adult dataset.

6. Reading A Complete Criterion for Value of Information in Soluble Influence Diagrams

The homomorphisms seem useful for transforming CIDs. I think I might be able to use them for proving which families of diagrams are incentive-ID, but not generally ID for Cft effects.

It cites r63 and r66! So those could be good directions to go.

References

1. Tom Everitt, Ryan Carey, Eric Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective, 2021.