CRAB and LEWIS Extensions

Ved Sriraman

December 2024

1 Introduction

This document contains extensions of proofs from the CRAB and LEWIS papers (Galhotra et. al) [5][2] .

2 Extending Section 3 of CRAB

2.1 Extension of Proposition 3.3

We first focus on the case of no external information. We will explore how to deal with uncertainty in multiple edges in a causal graph, and how that affects our bounds on the target distribution and thus the fairness bounds.

2.1.1 Preliminaries

Recall that we define fairness as such:

$$\mathsf{F}(\Omega) = \frac{1}{|\mathbf{A}|} \sum_{\mathbf{a} \in Dom(\mathbf{A})} \Pr^{+}_{\Omega}(h(x) \mid s_1, \mathbf{a}) - \Pr^{+}_{\Omega}(h(x) \mid s_0, \mathbf{a})$$

We will now examine two conditions in a causal graph for "perfect fairness" (PF).

PF1. If $(h(x) \perp\!\!\!\perp_d S \mid \mathbf{A})$, then $\Pr_{\Omega}(h(x) \mid S, \mathbf{A}) = \Pr_{\Omega}(h(x) \mid \mathbf{A}) \implies \mathsf{F}(\Omega) = 0$.

That is, if $h(x)$ is d-separated from S by $\mathbf{A}$, then we achieve perfect fairness as measured by $\mathsf{F}(\Omega)$.

PF2. If S is somehow isolated from $h(x)$, meaning that there is no path between S and $h(x)$, then $\Pr_{\Omega}(h(x) \mid S, \mathbf{A}) = \Pr_{\Omega}(h(x) \mid \mathbf{A}) \implies \mathsf{F}(\Omega) = 0$.

That is, knowing anything about S tells us absolutely nothing about Y, so we achieve perfect fairness as measured by $\mathsf{F}(\Omega)$.

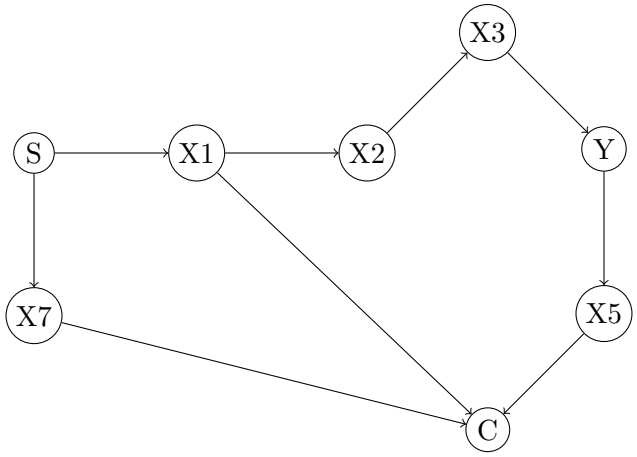
We will now apply these conditions to prove a statement about the conditional independence requirement in Proposition 3.3.

2.1.2 Uncertainty in only edge

Consider the causal graph $\mathcal{G} = (E, V)$ below that represents the true underlying model.

Here, we assume that all $x \in X$ are used to compute the target variable $h(x)$. Suppose that in a real-world application, we are unsure about the uncertainty of some causal link $X_i \to C$, and we can estimate from observational data that this exists in our causal graph with probability p. If $\mathcal{G}' = (E, V)$, with edges and nodes identical to the original causal graph $\mathcal{G}$, then we denote this uncertain edge as e_p. This represents our belief that the edge e_p exists with probability p. Note that, if $p = 1$, the edge is no longer uncertain, and $\mathcal{G}' = \mathcal{G}$. Let us also define that for the modified graph $\mathcal{G}'$,

$$E_{\mathbf{U}} = \big\{ e = (x_1, x_2) \in E \mid x_1 \in \mathbf{U} \text{ and } x_2 = C \big\}$$



That is, $E_{\mathbf{U}}$ is the set of all edges having a source in the set $\mathbf{U}$ from the original graph $\mathcal{G}$ and the sink as C itself. We want to show that if $e_p \notin E_{\mathbf{U}}$, $(\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{A}, \mathbf{U})$ still holds true. We will assume that neither C nor Y are parents of any other nodes in the causal graph.

Lemma 2.1. *Let $\mathcal{G}$ be the original causal graph with no uncertain edges and let $\mathcal{G}'$ be the modified causal graph with uncertain edge e_p. The independence condition $(\boldsymbol{X} \perp\!\!\!\perp C \mid S, \boldsymbol{A}, \boldsymbol{U})$ still holds in $\mathcal{G}'$ if we use the same $\mathbf{U}$ that satisfied the independence condition in $\mathcal{G}$, and the fairness query on the modified graph $\mathcal{G}'$ will be bounded by the fairness query on the original graph $\mathcal{G}$.*

Proof. All edges in $E_{\mathbf{U}}$ exist with certainty. Since $\mathbf{U}$ still forms the Markov boundary of C, conditioning on $\mathbf{U}$ will d-separate any other node in X from C. Thus, $(X \perp\!\!\!\perp_d C \mid S, \mathbf{A}, \mathbf{U}) \implies (\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{A}, \mathbf{U})$. □

Now, we can leverage this Lemma to derive improved bounds on the fairness query. WLOG, assume $p \in [0, \frac{1}{2}]$.

- If w.p. p we have a graph $\mathcal{G}'$ that satisfies *PF1* or *PF2* $\implies \mathsf{F}(\Omega) = 0$, we get the following two cases:

Case 1. if, w.p. $1-p$, we have a graph $\mathcal{G}'$ that satisfies *PF1* or *PF2* $\implies \mathsf{F}(\Omega) = 0$,

Case 2. else, w.p. $1-p$, we have a graph in which $U, S, \mathbf{A}$ are unchanged, yet neither *PF1* nor *PF2* are satisfied $\implies \mathsf{F}(\Omega) \leq (1-p) \cdot \mathit{CUB}$

- Else, w.p. p, we have a graph in which $U, S, \mathbf{A}$ are unchanged, yet neither *PF1* nor *PF2* are satisfied $\implies \mathsf{F}(\Omega) \leq p \cdot \mathit{CUB}$

Case 3. if, w.p. $1-p$, we have a graph $\mathcal{G}'$ that satisfies *PF1* or *PF2* $\implies \mathsf{F}(\Omega) = 0$,

Case 4. else w.p. $1-p$ we have a graph in which $\mathbf{U}, S, \mathbf{A}$ are unchanged, yet neither *PF1* nor *PF2* are satisfied $\implies \mathsf{F}(\Omega) \leq (1-p) \cdot \mathit{CUB}$

Case 1 exhibits perfect fairness because we obtain $\mathsf{F}(\Omega) = 0$ w.p. 1. Case 4 experiences no change in the fairness query as compared to the graph $\mathcal{G}$. Cases 2 and 3 achieve improved fairness bounds of $\mathsf{F}(\Omega) \leq (1-p) \cdot \mathit{CUB}$ and $\mathsf{F}(\Omega) \leq p \cdot \mathit{CUB}$, respectively.

Thus, by Lemma 2.1, we do not have to consider any uncertainty in edges that do not change the Markov boundary of C, i.e., a valid set U. We can shift our focus to the case where there exist uncertain edges in the set $E_{\mathbf{U}}$. For the singular case, take the example of one uncertain edge $e_p \in E_{\mathbf{U}}$ with source u_1. Let $\mathbf{U}' = \mathbf{U} \setminus \{u_1\}$. With a naive expected-case analysis, we can form the following consistent upper bound (CUB):

$$\mathit{CUB} = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A})} \left(p \cdot \left(\max_{\mathbf{u} \in \mathrm{Dom}(\mathbf{U})} \Pr^{+}_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}) - \min_{\mathbf{u} \in \mathrm{Dom}(\mathbf{U})} \Pr^{+}_{\Delta}(h(x) \mid s_0, \mathbf{a}, \mathbf{u}) \right) \right)$$

$$+\frac{1}{|A|}\sum_{\mathbf{a}\in\mathrm{Dom}(\mathbf{A})}\left((1-p)\cdot\left(\max_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr^{+}_{\Delta}(h(x)\mid s_1,\mathbf{a},\mathbf{u}')-\min_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr^{+}_{\Delta}(h(x)\mid s_0,\mathbf{a},\mathbf{u}')\right)\right)$$

We will prove an upper bound for each $\mathbf{a}\in\mathbf{A}$ in the definition of $\mathsf{F}_{h,\mathbf{A}}(\Omega)$, which extends to showing a lower bound and providing the desired result above.

Proof.

$$\begin{aligned}
\Pr_\Omega(h(x)\mid s_1,\mathbf{a}) &= p\cdot\sum_{\mathbf{u}\in\mathrm{Dom}(\mathbf{U})}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u})\cdot\Pr_\Omega(\mathbf{u}\mid s_1,\mathbf{a})\\
&\quad+(1-p)\cdot\sum_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u}')\cdot\Pr_\Omega(\mathbf{u}'\mid s_1,\mathbf{a})\\
&= p\cdot\sum_{\mathbf{u}\in\mathrm{Dom}(\mathbf{U})}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u},C=1)\cdot\Pr_\Omega(\mathbf{u}\mid s_1,\mathbf{a})\\
&\quad+(1-p)\cdot\sum_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u}',C=1)\cdot\Pr_\Omega(\mathbf{u}'\mid s_1,\mathbf{a})\\
&\le p\cdot\sum_{\mathbf{u}\in\mathrm{Dom}(\mathbf{U})}\max_{u\in\mathrm{Dom}(\mathbf{U})}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u},C=1)\cdot\Pr_\Omega(\mathbf{u}\mid s_1,\mathbf{a})\\
&\quad+(1-p)\cdot\sum_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\max_{u'\in\mathrm{Dom}(\mathbf{U}')}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u}',C=1)\cdot\Pr_\Omega(\mathbf{u}'\mid s_1,\mathbf{a})\\
&= p\cdot\max_{u\in\mathrm{Dom}(\mathbf{U})}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u},C=1)\sum_{\mathbf{u}\in\mathrm{Dom}(\mathbf{U})}\Pr_\Omega(\mathbf{u}\mid s_1,\mathbf{a})\\
&\quad+(1-p)\cdot\max_{u'\in\mathrm{Dom}(\mathbf{U}')}\Pr_\Omega(h(x)\mid s_1,\mathbf{a},\mathbf{u}',C=1)\cdot\sum_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr_\Omega(\mathbf{u}'\mid s_1,\mathbf{a})\\
&= p\cdot\Pr_\Delta(h(x)\mid s_1,\mathbf{a},\mathbf{u}^*)+(1-p)\cdot\Pr_\Delta(h(x)\mid s_1,\mathbf{a},\mathbf{u}'^*)
\end{aligned}$$

□

This naive approach requires us to iterate through possibilities of $\mathbf{u}$ and $\mathbf{u}'$. Instead, we can show a simplified approach, which allows us to form the following CUB. Letting β be a constant (defined below) between 0 and 2, we prove that:

$$CUB=\frac{1}{|A|}\sum_{\mathbf{a}\in\mathrm{Dom}(\mathbf{A})}\left((1+p\cdot\beta-p)\cdot\max_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr^{+}_{\Delta}(h(x)\mid s_1,\mathbf{a},\mathbf{u}')-(1+p\cdot\beta-p)\cdot\min_{\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')}\Pr^{+}_{\Delta}(h(x)\mid s_0,\mathbf{a},\mathbf{u}')\right)$$

Before we begin the proof, we will show some auxiliary work. We will consider $\mathbf{u}\in\mathrm{Dom}(\mathbf{U})$ and $\mathbf{u}'\in\mathrm{Dom}(\mathbf{U}')$, and abbreviate $\Pr_\Omega(X)$ as $\Pr(X)$ in order to show the following simplification:

$$\begin{aligned}
\sum_{\mathbf{u}}\Pr(h(x)\mid s_1,\mathbf{a},\mathbf{u},C=1)\cdot\Pr(\mathbf{u}\mid s_1,\mathbf{a}) &= \sum_{\mathbf{u}}\Pr(h(x)\mid s_1,\mathbf{a},\mathbf{u}'\wedge u_1,C=1)\cdot\Pr(\mathbf{u}'\wedge u_1\mid s_1,\mathbf{a})\\
&= \sum_{\mathbf{u}}\frac{\Pr(h(x),s_1,\mathbf{a},\mathbf{u}',u_1,C=1)}{\Pr(s_1,\mathbf{a},\mathbf{u}',u_1,C=1)}\cdot\frac{\Pr(\mathbf{u}',u_1,s_1,\mathbf{a})}{\Pr(s_1,\mathbf{a})}\\
&= \sum_{\mathbf{u}}\frac{\Pr(h(x),s_1,\mathbf{a},\mathbf{u}',u_1,C=1)}{\Pr(s_1,\mathbf{a},\mathbf{u}',u_1,C=1)}\cdot\frac{\Pr(\mathbf{u}',u_1,s_1,\mathbf{a})}{\Pr(s_1,\mathbf{a})}\cdot\frac{\Pr(\mathbf{u}',s_1,\mathbf{a},C=1)}{\Pr(\mathbf{u}',s_1,\mathbf{a},C=1)}\\
&= \sum_{\mathbf{u}}\Pr(h(x),u_1\mid s_1,\mathbf{a},\mathbf{u}',C=1)\cdot\Pr(\mathbf{u}',C=1\mid s_1,\mathbf{a})
\end{aligned}$$

$$
\begin{aligned}
&\cdot \frac{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a})}{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a}, C=1)} \\
&\leq \sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}', C=1 \mid s_1, \mathbf{a}) \\
&\quad \cdot \left\{ \frac{\Pr(\mathbf{u}', u_1=0, s_1, \mathbf{a})}{\Pr(\mathbf{u}', u_1=0, s_1, \mathbf{a}, C=1)} + \frac{\Pr(\mathbf{u}', u_1=1, s_1, \mathbf{a})}{\Pr(\mathbf{u}', u_1=1, s_1, \mathbf{a}, C=1)} \right\} \\
&\leq \sum_{\mathbf{u}'} \left[\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \right. \\
&\quad \left. \cdot \left\{ \frac{\Pr(\mathbf{u}', u_1=0, s_1, \mathbf{a})}{\Pr(\mathbf{u}', u_1=0, s_1, \mathbf{a}, C=1)} + \frac{\Pr(\mathbf{u}', u_1=1, s_1, \mathbf{a})}{\Pr(\mathbf{u}', u_1=1, s_1, \mathbf{a}, C=1)} \right\} \right] \\
&\quad \cdot \Pr(\mathbf{u}', C=1 \mid s_1, \mathbf{a}) \\
&= \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*, C=1) \\
&\quad \cdot \left\{ \frac{\Pr(\mathbf{u}'^*, u_1=0, s_1, \mathbf{a})}{\Pr(\mathbf{u}'^*, u_1=0, s_1, \mathbf{a}, C=1)} + \frac{\Pr(\mathbf{u}'^*, u_1=1, s_1, \mathbf{a})}{\Pr(\mathbf{u}'^*, u_1=1, s_1, \mathbf{a}, C=1)} \right\} \\
&\quad \cdot \sum_{\mathbf{u}'} \Pr(\mathbf{u}', C=1 \mid s_1, \mathbf{a}) \\
&= \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*, C=1) \cdot \beta \cdot \sum_{\mathbf{u}'} \Pr(\mathbf{u}', C=1 \mid s_1, \mathbf{a}) \\
&= \beta \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*, C=1) \cdot \Pr(C=1 \mid s_1, \mathbf{a}) \\
&= \beta \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*, C=1) \cdot (1 - \Pr(C=0 \mid s_1, \mathbf{a})) \\
&\leq \beta \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*, C=1) \\
&= \beta \cdot \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*)
\end{aligned}
$$

Now, we can provide an upper bound on the expected value of the fairness query given one uncertain edge as follows:

Proof.

$$
\begin{aligned}
\Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}) &= p \cdot \sum_{\mathbf{u} \in \mathrm{Dom}(\mathbf{U})} \Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}) \cdot \Pr_{\Omega}(\mathbf{u} \mid s_1, \mathbf{a}) \\
&\quad + (1-p) \cdot \sum_{\mathbf{u}' \in \mathrm{Dom}(\mathbf{U}')} \Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}') \cdot \Pr_{\Omega}(\mathbf{u}' \mid s_1, \mathbf{a}) \\
&= p \cdot \sum_{\mathbf{u} \in \mathrm{Dom}(\mathbf{U})} \Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr_{\Omega}(\mathbf{u} \mid s_1, \mathbf{a}) \\
&\quad + (1-p) \cdot \sum_{\mathbf{u}' \in \mathrm{Dom}(\mathbf{U}')} \Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr_{\Omega}(\mathbf{u}' \mid s_1, \mathbf{a}) \\
&\leq p \cdot \beta \cdot \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*) + (1-p) \cdot \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*) \\
&= (1 + p \cdot \beta - p) \cdot \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'^*)
\end{aligned}
$$

We can show the result above by applying similar logic to find the minimum. □

However, upon taking a closer look at the β factor in this second analysis, we note that this bound is impractical because it is not possible to calculate $\Pr(\mathbf{u}'^*, u_1=0, s_1, \mathbf{a})$ or $\Pr(\mathbf{u}'^*, u_1=1, s_1, \mathbf{a})$ from observational data under the influence of selection bias. Instead, we can make a slight adjustment to the first part of our previous analysis, and we will use the following result to upper bound $\Pr_{\Omega}(h(x) \mid s_1, \mathbf{a})$ itself in the worst case.

$$
\sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) = \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1, C=1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a})
$$

$$
\begin{aligned}
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \cdot \frac{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)}{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)} \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid s_1, u_1, \mathbf{u}', \mathbf{a})} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \sum_{\mathbf{u}'} \sum_{\mathbf{u_1}} \frac{\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid u_1, \mathbf{u}')} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\leq \sum_{\mathbf{u}'} \sum_{\mathbf{u_1}} \frac{\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\leq \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \sum_{\mathbf{u}'} \sum_{\mathbf{u_1}} \Big(\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&\quad \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \Big) \\
&= \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \sum_{\mathbf{u_1}} \Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&= \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&\leq \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}} \Pr(C=1, \mathbf{u}' \mid \mathbf{u})} \cdot \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \Pr(C=1 \mid s_1, \mathbf{a}) \\
&= \frac{\Pr(C=1 \mid s_1, \mathbf{a})}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}') \\
&= \alpha \cdot \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')
\end{aligned}
$$

Next, we will form the new CUB by considering the uncertainty in one edge. In the case of one uncertain edge, we will take the maximum of the case in which u_1 exists in the causal graph and the case in which u_1 does not exist.

$$
\begin{aligned}
\Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}) &\leq \max \left\{ \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}), \sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}) \right\} \\
&= \max \left\{ \alpha \cdot \max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'), \max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}') \right\} \\
&= \max \{\alpha, 1\} \cdot \max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')
\end{aligned}
$$

Thus, we have the following CUB:

$$
CUB = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A})} \left(\max \{\alpha, 1\} \cdot \max_{\mathbf{u}' \in \mathrm{Dom}(\mathbf{U}')} \Pr^{+}_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}') - \min \{\alpha, 1\} \cdot \min_{\mathbf{u}' \in \mathrm{Dom}(\mathbf{U}')} \Pr^{+}_{\Delta}(h(x) \mid s_0, \mathbf{a}, \mathbf{u}') \right)
$$

This approach to a worst-case analysis is feasible because we can estimate $\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})$ through observational data, and thus our analysis is tighter than the previous result. This is less intensive than calculating the conditional distribution we approximated over $\mathbf{u} \in \mathbf{U}$. In this analysis, the only approximation we have to make is on the conditional distribution of the target $h(x)$ by dropping the intersection with u_1 to reduce the complexity.

2.1.3 Uncertainty in multiple edges

With multiple uncertainties in causal links provided by nodes in $\mathbf{U}$, we must consider the maximum over all possible sets $\mathbf{U}$. If there are ℓ uncertain edges, then 2^ℓ observable sets of $\mathbf{U}$ are possible, all of which include the nodes with no uncertainty. Consider an example with two uncertain edges from u_1 and u_2, and $\mathbf{u}' \in \mathbf{U}' = \mathbf{U} \setminus \{u_1, u_2\}$. We will take the max over four separate cases.

$$\begin{aligned}\Pr_\Omega(h(x) \mid s_1, \mathbf{a}) \leq \max \Bigg\{ & \sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}), \\ & \sum_{\mathbf{u}' \wedge u_1} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1, C=1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a}), \\ & \sum_{\mathbf{u}' \wedge u_2} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_2, C=1) \cdot \Pr(\mathbf{u}' \wedge u_2 \mid s_1, \mathbf{a}), \\ & \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) \Bigg\}\end{aligned}$$

We will simplify each of the four cases below:

Case 1. $u_1, u_2 \notin \mathbf{U}$ We cite the proof of Proposition 3.1 from the CRAB paper, since $(\mathbf{X} \perp\!\!\!\perp C \,|\, S, \mathbf{A}, \mathbf{U})$ is satisfied.

$$\sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}) \leq \max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')$$

Case 2. $u_1 \in \mathbf{U}, u_2 \notin \mathbf{U}$ The result extends exactly from the work in Case 1. However, the step with the first inequality is less lossy because we only discard u_2 from the conditional joint distribution.

$$\begin{aligned}\sum_{\mathbf{u}' \wedge u_1} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1, C=1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a}) &= \sum_{\mathbf{u}' \wedge u_1} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', u_1, C=1) \cdot \Pr(\mathbf{u}', u_1 \mid s_1, \mathbf{a}) \\ &= \sum_{\mathbf{u}' \wedge u_1} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \\ &= \sum_{\mathbf{u}' \wedge u_1} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \\ &\quad \cdot \frac{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)}{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)} \\ &= \sum_{\mathbf{u}' \wedge u_1} \frac{\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid s_1, u_1, \mathbf{u}', \mathbf{a})} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\ &= \sum_{\mathbf{u}' \wedge u_1} \frac{\Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid u_1, \mathbf{u}')} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\ &\leq \frac{1}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \sum_{\mathbf{u}' \wedge u_1} \Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\ &\quad \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\ &= \frac{1}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a})\end{aligned}$$

$$
\begin{aligned}
&\sum_{u_1} \Pr(h(x), u_1 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&= \frac{1}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\quad \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&\leq \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \cdot \Pr(C=1 \mid s_1, \mathbf{a}) \\
&= \frac{\max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \cdot \Pr(C=1 \mid s_1, \mathbf{a})
\end{aligned}
$$

Case 3. $u_2 \in \mathbf{U}, u_1 \notin \mathbf{U}$

Similar reasoning to the previous case.

Case 4. $u_1, u_2 \in \mathbf{U}$

Note that ($\mathbf{X} \not\perp\!\!\!\perp C \,|\, S, \mathbf{A}, \mathbf{U}$').

$$
\begin{aligned}
\sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) &= \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge \{u_1, u_2\}, C=1) \cdot \Pr(\mathbf{u}' \wedge \{u_1, u_2\} \mid s_1, \mathbf{a}) \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, u_2, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, u_2, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, u_2, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), s_1, \mathbf{a}, \mathbf{u}', u_1, u_2, C=1)}{\Pr(s_1, \mathbf{a}, \mathbf{u}', u_1, u_2, C=1)} \cdot \frac{\Pr(\mathbf{u}', u_1, u_2, s_1, \mathbf{a})}{\Pr(s_1, \mathbf{a})} \cdot \frac{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)}{\Pr(\mathbf{u}', s_1, \mathbf{a}, C=1)} \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), u_1, u_2 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid s_1, u_1, u_2, \mathbf{u}', \mathbf{a})} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \sum_{\mathbf{u}} \frac{\Pr(h(x), u_1, u_2 \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\Pr(C=1 \mid u_1, u_2, \mathbf{u}')} \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\leq \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid u_1, u_2, \mathbf{u}')} \sum_{\mathbf{u}} \Pr(h(x), u_1, u_2 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&\quad \cdot \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid u_1, u_2, \mathbf{u}')} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\quad \sum_{u_1, u_2} \Pr(h(x), u_1, u_2 \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&= \frac{1}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&\quad \cdot \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \\
&\leq \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \sum_{\mathbf{u}'} \Pr(C=1, \mathbf{u}' \mid s_1, \mathbf{a}) \\
&= \frac{\max_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1)}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \Pr(C=1 \mid s_1, \mathbf{a}) \\
&= \frac{\max_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \Pr(C=1 \mid s_1, \mathbf{a})
\end{aligned}
$$

We can now write the final bound.

$$
\begin{aligned}
\Pr_\Omega(h(x) \mid s_1, \mathbf{a}) \leq \max \Bigg\{ & \sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_1} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1, C=1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_2} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_2, C=1) \cdot \Pr(\mathbf{u}' \wedge u_2 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) \Bigg\} \\
= \max \Bigg\{ & \max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}'), \\
& \frac{\max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')} \cdot \Pr(C=1 \mid s_1, \mathbf{a}), \\
& \frac{\max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')}{\min_{\mathbf{u}' \wedge u_2} \Pr(C=1 \mid u_2, \mathbf{u}')} \cdot \Pr(C=1 \mid s_1, \mathbf{a}), \\
& \frac{\max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \cdot \Pr(C=1 \mid s_1, \mathbf{a}) \Bigg\} \\
= \max \Bigg\{ & 1, \frac{\Pr(C=1 \mid s_1, \mathbf{a})}{\min_{\mathbf{u}' \wedge u_1} \Pr(C=1 \mid u_1, \mathbf{u}')}, \frac{\Pr(C=1 \mid s_1, \mathbf{a})}{\min_{\mathbf{u}' \wedge u_2} \Pr(C=1 \mid u_2, \mathbf{u}')}, \frac{\Pr(C=1 \mid s_1, \mathbf{a})}{\min_{\mathbf{u}} \Pr(C=1 \mid \mathbf{u})} \Bigg\} \\
& \cdot \max_{\mathbf{u}'} \Pr_\Delta(h(x) \mid s_1, \mathbf{a}, \mathbf{u}')
\end{aligned}
$$

This type of reasoning easily extends beyond two edges, but we must be careful about computability, which is discussed in the next section.

2.1.4 Computability of Our Bounds

Note that the two types of noise we are dealing with here come from (1) determining what the selection variable C is, and (2), determining what the parents of C are. Thus, to quantify the overall complexity of this bound when extending to uncertainty in multiple edges, we can separate the calculation into two parts:

1. We must first choose a set of ℓ uncertain edges, such that as ℓ increases, the number of terms in the max grows exponentially as 2^ℓ. This takes 2^n time if every edge is considered. In real-world applications, since we are usually certain (with high probability) that only a certain number of k features are non-latent variables in the causal graph, this complexity effectively reduces to $\binom{n}{k}$. Note that this step can be greedily approximated using a decision tree searching algorithm, which reduces the term to $n \times k$. This involves repeatedly choosing one of n edges that maximizes the bound in the case of uncertainty in one edge.

2. Choosing k chosen variables (using our bound for k uncertain edges) would generate a $\mathcal{O}(2^k)$ complexity term. By treating only certain "important" nodes as uncertain, and the rest part of $\mathbf{U}'$, we arrive at a trade-off between the complexity of our bound, which will reduce by a factor of 2 for each one left out of consideration, and the accuracy of our bound, which decreases with more variables left out. We can explore this trade-off using the range of possible bounds that we have, which differ in the number of uncertain edges considered.

3. Then, for each set of edges that constitute the Markov boundary of C in the causal graph, we must evaluate the α-like quantity based on the uncertain edges we have chosen, which serves as the current $\text{Dom}(Pa(C))$. In the case of binary values for nodes in the causal graph, iterating over all possible assignments to each of the k variables would generate a worst-case complexity of $\mathcal{O}(2^k)$. To further reduce the complexity of this bound, we could greedily choose only important nodes from $\mathbf{U} \setminus \mathbf{U}'$ that we believe contribute most to our

bound, and leave out others. This would reduce our complexity by a factor of 2 for each one left out of consideration. However, this leads to a more lossy bound.

2.2 Extension of Proposition 3.4

Now we establish conditions under which $\mathsf{F}(\Omega)$ can be estimated or tightly bounded from biased data when we have access to external data sources that reveal varying levels of information about the target population. When we have access to auxiliary statistics, evaluating each possible causal graph is straightforward. We obtain the following bounds when dealing with uncertainty in multiple edges.

$$\begin{aligned}
\Pr_{\Omega}(h(x) \mid s_1, \mathbf{a}) \leq \max \Bigg\{ & \sum_{\mathbf{u}'} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}', C=1) \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_1} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1, C=1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_2} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_2, C=1) \cdot \Pr(\mathbf{u}' \wedge u_2 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}} \Pr(h(x) \mid s_1, \mathbf{a}, \mathbf{u}, C=1) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) \Bigg\} \\
= \max \Bigg\{ & \sum_{\mathbf{u}'} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}') \cdot \Pr(\mathbf{u}' \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_1} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_1) \cdot \Pr(\mathbf{u}' \wedge u_1 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}' \wedge u_2} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}' \wedge u_2) \cdot \Pr(\mathbf{u}' \wedge u_2 \mid s_1, \mathbf{a}), \\
& \sum_{\mathbf{u}} \Pr_{\Delta}(h(x) \mid s_1, \mathbf{a}, \mathbf{u}) \cdot \Pr(\mathbf{u} \mid s_1, \mathbf{a}) \Bigg\}
\end{aligned}$$

This type of reasoning again extends beyond two edges.

3 Properties of Explanation Scores

Recent work by Galhotra et al. [2] introduces a novel framework for Explainable AI (XAI) centered on probabilistic contrastive counterfactuals, inspired by human cognitive processes. It addresses limitations of existing XAI methods, which often rely on associational analysis rather than causal relationships. The proposed system, LEWIS, provides local, global, and contextual explanations for decisions made by black-box algorithms while generating actionable recourse for negatively affected individuals. Unlike traditional methods, LEWIS incorporates causal reasoning to identify both direct and indirect influences of attributes, ensuring its recommendations are actionable in real-world scenarios. The framework is model-agnostic, requiring only input-output data, and can work across varying levels of user understanding of causal models. However, LEWIS assumes access to population statistics in its interventional calculations, which diminishes its applicability to real-world settings.

In this section, we aim to improve the applicability of LEWIS by studying the properties of the scores used for generating counterfactual explanations and establish conditions under which they can be bounded or estimated from historical data (Section X), even in the presence of selection bias. Specifically, we first calculate exact scores for binary attribute values in the presence of selection bias. Then, we turn our attention to a more general setting, in which we leverage applicable conditions to achieve tight upper bounds.

Symbol	Meaning
X, Y, Z	attributes (variables)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	sets of attributes
$Dom(X), Dom(\mathbf{X})$	their domains
$x \in Dom(X)$	an attribute value
$\mathbf{x} \in Dom(\mathbf{X})$	a tuple of attribute values
$\mathbf{k} \in Dom(\mathbf{K})$	a tuple of context attribute values
$\mathcal{G}$	causal diagram
$\langle M, \Pr(\mathbf{u})\rangle$	probabilistic causal model
$O_{\mathbf{X}\leftarrow\mathbf{x}}$	potential outcome
$\Pr_\Omega(\mathbf{V}=\mathbf{v}), \Pr(\mathbf{v})$	true joint probability distribution
$\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}})$	abbreviates $\Pr_\Omega(O_{\mathbf{X}\leftarrow\mathbf{x}}=o)$
$\Pr_\Delta(\mathbf{V}=\mathbf{v}), \Pr_\Delta(\mathbf{v})$	joint distribution over biased data

Table 1: **Notation used in this section.**

3.1 Preliminaries

At the heart of [2] are *probabilistic contrastive counterfactuals* of the following form:

> "For individual(s) with attribute(s) <actual-value> for whom an algorithm made the decision <actual-outcome>, the decision would have been <foil-outcome> with *probability* <score> had the attribute been <counterfactual-value>." (1)

The notation we use in this paper is summarized in Table 1. We denote variables by uppercase letters, X, Y, Z, V; their values with lowercase letters, x, y, z, v; and sets of variables or values using boldface ($\mathbf{X}$ or $\mathbf{x}$). The domain of a variable X is $Dom(X)$, and the domain of a set of variables is $Dom(\mathbf{X}) = \prod_{X\in\mathbf{X}} Dom(X)$. All domains are discrete and finite; continuous domains are assumed to be binned. We use $\Pr(\mathbf{x})$ to represent a joint probability distribution $\Pr(\mathbf{X}=\mathbf{x})$. The basic semantic framework of our proposal rests on probabilistic causal models [3], which we review next.

Probabilistic causal models. A *probabilistic causal model* (PCM) is a tuple $\langle M, \Pr(\mathbf{u})\rangle$, where $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}\rangle$ is a *causal model* consisting of a set of *observable or endogenous* variables $\mathbf{V}$ and a set of *background or exogenous* variables $\mathbf{U}$ that are outside of the model, and $\mathbf{F} = (F_X)_{X\in\mathbf{V}}$ is a set of *structural equations* of the form $F_X : Dom(\mathbf{Pa_V}(X)) \times Dom(\mathbf{Pa_U}(X)) \to Dom(X)$, where $\mathbf{Pa_U}(X) \subseteq \mathbf{U}$ and $\mathbf{Pa_V}(X) \subseteq \mathbf{V} - \{X\}$ are called *exogenous parents* and *endogenous parents* of X, respectively. The values of $\mathbf{U}$ are drawn from the distribution $\Pr(\mathbf{u})$. A PCM $\langle M, \Pr(\mathbf{u})\rangle$ can be represented as a directed graph $\mathcal{G} = \langle \mathbf{V}, \mathbf{E}\rangle$, called a *causal diagram*, where each node represents a variable, and there are directed edges from the elements of $\mathbf{Pa_U}(X) \cup \mathbf{Pa_V}(X)$ to X. We say a variable Z is a *descendant* of another variable X if Z is *caused* (either *directly* or *indirectly*) by X, i.e., if there is a directed edge or path from X to Z in G; otherwise, we say that Z is a *non-descendant* of X.

Interventions and potential outcomes. An *intervention* or an *action* on a set of variables $\mathbf{X} \subseteq \mathbf{V}$, denoted $\mathbf{X} \leftarrow \mathbf{x}$, is an operation that *modifies* the underlying causal model by replacing the structural equations associated with $\mathbf{X}$ with a constant $\mathbf{x} \in Dom(\mathbf{X})$. The *potential outcome* of a variable Y after the intervention $\mathbf{X} \leftarrow \mathbf{x}$ in a context $\mathbf{u} \in Dom(\mathbf{U})$, denoted $Y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u})$, is the *solution* to Y in the modified set of structural equations. Potential outcomes satisfy the following *consistency rule* used in derivations presented in Section **??**.

$$\mathbf{X}(\mathbf{u}) = \mathbf{x} \implies Y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) = y \tag{2}$$

This rule states that in contexts where $\mathbf{X} = \mathbf{x}$, the outcome is invariant to the intervention $\mathbf{X} \leftarrow \mathbf{x}$. For example, changing the income-level of applicants to high does not change the loan decisions for those who already had high-incomes before the intervention.

The distribution $\Pr(\mathbf{u})$ induces a probability distribution over endogenous variables and potential outcomes. Using PCMs, one can express *counterfactual queries* of the form $\Pr(Y_{\mathbf{X}\leftarrow\mathbf{x}} = y \mid \mathbf{k})$, or simply $\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k})$; this reads as "For contexts with attributes $\mathbf{k}$, what is the probability that we would observe $Y = y$ had $\mathbf{X}$ been $\mathbf{x}$?" and is given by the following expression:

$$\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k}) = \sum_{\mathbf{u}} \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}))\ \Pr(\mathbf{u} \mid \mathbf{k}) \tag{3}$$

Equation (3) readily suggests Pearl's three-step procedure for answering counterfactual queries [3][Chapter 7]: (1) update $\Pr(\mathbf{u})$ to obtain $\Pr(\mathbf{u} \mid \mathbf{k})$ (*abduction*), (2) modify the causal model to reflect the intervention $\mathbf{X} \leftarrow \mathbf{x}$ (*action*), and (3) evaluate the RHS of (3) using the index function $\Pr(Y_{\mathbf{X} \leftarrow \mathbf{x}}(\mathbf{u}) = y)$ (*prediction).* However, performing this procedure requires the underlying PCM to be fully observed, i.e, the distribution $\Pr(\mathbf{u})$ and the underlying structural equations must be known, which is an impractical requirement. In this paper, we assume that only background knowledge of the underlying causal diagram is available, but exogenous variables and structural equations are unknown.

The do-operator. For causal diagrams, Pearl defined the do-operator as a graphical operation that gives semantics to *interventional queries* of the form "What is the probability that we would observe $Y = y$ (at population-level) had $\mathbf{X}$ been $\mathbf{x}$?", denoted $\Pr(\mathbf{y} \mid \texttt{do}(\mathbf{x}))$. Further, he proved a set of necessary and sufficient conditions under which interventional queries can be answered using historical data. A sufficient condition is the backdoor-criterion, [1] which states that if there exists a set of variables $\mathbf{C}$ that satisfy a graphical condition relative to $\mathbf{X}$ and Y in the causal diagram G, the following holds (see [3][Chapter 3] for details):

$$\Pr(\mathbf{y} \mid \texttt{do}(\mathbf{x})) = \sum_{\mathbf{c} \in Dom(\mathbf{C})} \Pr(\mathbf{y} \mid \mathbf{c}, \mathbf{x}) \Pr(\mathbf{c}) \tag{4}$$

In contrast to (3), notice that the RHS of (4) is expressed in terms of observed probabilities and can be estimated from historical data using existing statistical and ML algorithms.

Counterfactuals vs. interventional queries. The do-operator is a population-level operator, meaning it can only express queries about the effect of an intervention at population level; in contrast, counterfactuals can express queries about the effect of an intervention on a sub-population or an individual. Therefore, every interventional query can be expressed in terms of counterfactuals, but not vice versa (see [4][Chapter 4] for more details). For instance, $\Pr(y \mid \texttt{do}(\mathbf{x})) = \Pr(y_{\mathbf{X} \leftarrow \mathbf{x}})$; however, the counterfactual query $\Pr(y_{\mathbf{X} \leftarrow \mathbf{x}} \mid \mathbf{x}', y')$, which asks about the effect of the intervention $\mathbf{X} \leftarrow \mathbf{x}$ on a sub-population with attributes $\mathbf{x}'$ and y', cannot be expressed in terms of the do-operator. Note that the probabilistic contrastive counterfactual statements in (1), used throughout this paper to explain a black-box decision-making system concerned with the effect of interventions at sub-population and individual levels, cannot be expressed using the do-operator and therefore cannot be assessed in general when the underlying probabilistic causal models are not fully observed. Nevertheless, in Sections 3.3 and 3.4 we establish conditions under which these counterfactuals can be estimated or bounded using biased data.

3.2 Explanation Scores

We are given a decision-making algorithm $f : Dom(\mathbf{I}) \rightarrow Dom(O)$, where $\mathbf{I}$ is set of input attributes (a.k.a. features for ML algorithms) and O is a binary attribute, where $O = o$ denotes the positive decision (loan approved) and $O = o'$ denotes the negative decision (loan denied). Let us assume we are given a PCM $\langle M, \Pr(\mathbf{u}) \rangle$ with a corresponding causal diagram $\mathcal{G}$ (this assumption will be relaxed in Sections 3.3 and 3.4) such that $\mathbf{I} \subseteq \mathbf{V}$, i.e., the inputs of f are a subset of the observed attributes. Consider an attribute $X \in \mathbf{V}$ and a pair of attribute values $x, x' \in Dom(X)$. We quantify the influence of the attribute value x relative to a baseline x' on decisions made by an algorithm using the following scores, herein referred to as *explanation scores*; (we implicitly assume an order $x > x'$).

Definition 3.1 (Explanation Scores)**.** *Given a PCM $\langle M, \Pr(\mathbf{u}) \rangle$ and an algorithm $f : Dom(\mathbf{X}) \rightarrow Dom(O)$, a variable $X \in \mathbf{V}$, and a pair of attribute values $x, x' \in Dom(X)$, we quantify the influence of x relative to x' on the algorithm's decisions in the context $\mathbf{k} \in Dom(\mathbf{K})$, where $\mathbf{K} \subseteq \mathbf{V} - \{X, O\}$, using the following measures:*

- *The necessity score:*

$$\textsc{Nec}_x^{x'}(\mathbf{k}) \stackrel{def}{=} \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k}) \tag{5}$$

- *The sufficiency score:*

$$\textsc{Suf}_x^{x'}(\mathbf{k}) \stackrel{def}{=} \Pr(o_{X \leftarrow x} \mid x', o', \mathbf{k}) \tag{6}$$

[1] Since it is not needed in our work, we do not discuss the graph-theoretic notion of backdoor-criterion.

- *The necessity and sufficiency score:*

$$\textsc{NeSuf}_x^{x'}(\mathbf{k}) \stackrel{def}{=} \Pr(o_{X\leftarrow x}, o'_{X\leftarrow x'} \mid \mathbf{k}), \tag{7}$$

where the distribution $\Pr(o_{X\leftarrow x})$ *is well-defined and can be computed from the algorithm* $f(\mathbf{I})$.[2]

For simplicity of notation, we drop x' from $\textsc{Nec}_x^{x'}, \textsc{Suf}_x^{x'}$ and $\textsc{NeSuf}_x^{x'}$ whenever it is clear from the context. The necessity score in (5) formalizes the probabilistic contrastive counterfactual in (1), where <actual-value> and <counterfactual-value> are respectively $\mathbf{k} \cup x$ and $\mathbf{k} \cup x'$, and <actual-decision> and <foil-decision> are respectively positive decision o and negative decision o'. This reads as "What is the probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *negative* instead of *positive* had X been x' instead of x?" In other words, $\textsc{Nec}_X(.)$ measures the algorithm's percentage of positive decisions that are *attributable to* or *due to* the attribute value x. The sufficiency score in (6) is the dual of the necessity score; it reads as "What would be the probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *positive* instead of *negative* had X been x instead of x'?" Finally, the necessity and sufficiency score in (7) establishes a balance between necessary and sufficiency; it measures the probability that the algorithm responds in both ways. Hence, it can be used to measure the general explanatory power of an attribute. In Section 5, we show that the necessary and sufficiency score is non-zero iff X causally influences the algorithm's decisions. (Note that the explanation scores are well-defined for a set of attributes.)

3.3 Computing Exact Explanation Scores

Recall from Section 3.2 that if the underlying PCM is fully specified, i.e., the structural equations and the exogenous variables are observed, then counterfactual queries, and hence the explanation scores, can be computed via Equation 1. However, in many applications, PCMs are not fully observed, and one must estimate explanation scores from data. First, we show how to calculate explanation scores exactly, computed for a set of attributes $\mathbf{X}$.

For the sake of simplicity, we assume binary attribute values in this setting. A key assumption that we introduce, motivated by Galhotra and Halpern [1], is the following interventional independence:

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \mathbf{k}) \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \mathbf{k}) \tag{1}$$

This condition is applicable to settings in which conditioning on a particular attribute value and determining its effect on a positive outcome has no effect on whether a negative outcome would be achieved had we switched to a different attribute value.

Proposition 3.1. *Given a PCM* $\langle M', \Pr_\Delta(\mathbf{u})\rangle$ *with a corresponding causal DAG* $\mathcal{G}$ *which represents the underlying biased data collection process, an algorithm* $f : Dom(\mathbf{I}) \to Dom(O)$*, and a set of attributes* $\mathbf{X} \subseteq \mathbf{V} - \{O\}$ *with two sets of binary attribute values* $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$*, if* $\mathbf{K}$ *consists of non-descendants of* $\mathbf{I}$ *in* $\mathcal{G}$ *and if either of the following conditions hold:*

1. *There is no parent of* $\mathbf{O}$ *or* $\mathbf{X}$ *other than* $\mathbf{K}$*, and* C *is not a child of* $\mathbf{O}$ *or* $\mathbf{X}$*.*
2. *There is no parent of* C *other than* $\mathbf{K}$*.*

then the explanation scores can be calculated exactly as follows for any set of variables $\mathbf{K}$ *such that* $(\mathbf{O}, \mathbf{X} \perp\!\!\!\perp C \mid \mathbf{K})$*:*

$$\textsc{Nec}_{\mathbf{x}}(\mathbf{k}) = \frac{\big(1 - \Pr_\Delta(o', \mathbf{x} \mid \mathbf{k})\big) \cdot \Pr_\Delta(o' \mid \mathit{do}(\mathbf{x}'), \mathbf{k}) - \Pr_\Delta(o' \mid \mathbf{k}) + \Pr_\Delta(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(1 - \Pr_\Delta(o, \mathbf{x}' \mid \mathbf{k})\big)}{\Pr_\Delta(o, \mathbf{x} \mid \mathbf{k})} \tag{2}$$

$$\textsc{Suf}_{\mathbf{x}}(\mathbf{k}) = \frac{\big(\Pr_\Delta(o \mid \mathbf{x}', \mathbf{k}) - 1\big) \cdot \Pr_\Delta(o' \mid \mathit{do}(\mathbf{x}), \mathbf{k}) + \Pr_\Delta(o' \mid \mathbf{k}) - \Pr_\Delta(o \mid \mathbf{x}', \mathbf{k}) \cdot \Pr_\Delta(o', \mathbf{x} \mid \mathbf{k})}{\Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k})} \tag{3}$$

$$\textsc{NeSuf}_{\mathbf{x}}(\mathbf{k}) = \Pr_\Delta(o, \mathbf{x} \mid \mathbf{k}) \cdot \textsc{Nec}_{\mathbf{x}}(\mathbf{k}) + \Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k}) \cdot \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \tag{4}$$

[2]For deterministic $f(\mathbf{I})$, $\Pr(o_{X\leftarrow x}) = \sum_{\mathbf{i}\in Dom(\mathbf{I})} \mathbb{1}_{\{f(\mathbf{i})=o\}} \Pr(\mathbf{I}_{X\leftarrow x} = \mathbf{i})$, where $\mathbb{1}_{\{f(\mathbf{i})=o\}}$ is an indicator function.

Proof. We prove the bounds for (2); (3) is proved similarly, and (4) is derived from Proposition 5.1 in Section 5. The following equations are obtained from the law of total probability:

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) \tag{5}$$

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, o_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) \tag{6}$$

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}' \mid \mathbf{k}) \tag{7}$$

By rearranging (5) and (6), we obtain the following equality:

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, o_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) \tag{8}$$

The following bounds for the LHS of (8) are obtained using the interventional independence assumption in (1):

$$\begin{aligned}
\text{LHS} &= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) \cdot \frac{\Pr(\mathbf{x} \mid \mathbf{k})}{\Pr(\mathbf{x} \mid \mathbf{k})} \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \mathbf{k}) \cdot \Pr(\mathbf{x} \mid \mathbf{k}) \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \mathbf{k}) \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \mathbf{k}) \cdot \Pr(\mathbf{x} \mid \mathbf{k}) \\
&\qquad \texttt{(obtained from Eq.(1))} \\
&= \Big(\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}' \mid \mathbf{k})\Big) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \mathbf{k}) \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \mathbf{k}) \\
&\qquad \texttt{(obtained from Eq.(7))} \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}' \mid \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{x} \mid \mathbf{k}) \\
&\quad + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \mathbf{k}) \cdot \Big(\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}' \mid \mathbf{k})\Big) \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o', \mathbf{x}' \mid \mathbf{k}) - \Pr(o', \mathbf{x} \mid \mathbf{k}) + \Pr(o' \mid \mathbf{x}, \mathbf{k}) \cdot \Big(\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o, \mathbf{x}' \mid \mathbf{k})\Big) \\
&\qquad \texttt{(obtained from the consistency rule)} \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o' \mid \mathbf{k}) + \Pr(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o, \mathbf{x}' \mid \mathbf{k})\big) \\
&= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o' \mid \mathbf{k}) + \Pr(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(1 - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o, \mathbf{x}' \mid \mathbf{k})\big) \\
&= \big(1 - \Pr(o', \mathbf{x} \mid \mathbf{k})\big) \cdot \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) - \Pr(o' \mid \mathbf{k}) + \Pr(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(1 - \Pr(o, \mathbf{x}' \mid \mathbf{k})\big) \\
&= \big(1 - \Pr(o', \mathbf{x} \mid \mathbf{k})\big) \cdot \Pr(o' \mid \texttt{do}(\mathbf{x}'), \mathbf{k}) - \Pr(o' \mid \mathbf{k}) + \Pr(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(1 - \Pr(o, \mathbf{x}' \mid \mathbf{k})\big) \\
&\qquad \texttt{(obtained from the backdoor criterion)} \\
&= \big(1 - \Pr(o', \mathbf{x} \mid \mathbf{k}, C = 1)\big) \cdot \Pr(o' \mid \texttt{do}(\mathbf{x}'), \mathbf{k}, C = 1) - \Pr(o' \mid \mathbf{k}, C = 1) \\
&\quad + \Pr(o' \mid \mathbf{x}, \mathbf{k}, C = 1) \cdot \big(1 - \Pr(o, \mathbf{x}' \mid \mathbf{k}, C = 1)\big) \\
&\qquad \texttt{(obtained from the conditional independence assumption)} \\
&= \big(1 - \Pr_\Delta(o', \mathbf{x} \mid \mathbf{k})\big) \cdot \Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), \mathbf{k}) - \Pr_\Delta(o' \mid \mathbf{k}) \\
&\quad + \Pr_\Delta(o' \mid \mathbf{x}, \mathbf{k}) \cdot \big(1 - \Pr_\Delta(o, \mathbf{x}' \mid \mathbf{k})\big)
\end{aligned}$$

Dividing through by $\Pr(o, \mathbf{x} \mid \mathbf{k})$, we obtain the final result. □

3.4 Computing Explanation Scores: General case

In the general case, the assumption that $\text{Pa}(\mathbf{O})$, $\mathbf{O}$, and $\mathbf{X}$ are all disjoint from $\text{Pa}(C)$ constitutes one set of sufficient conditions for calculating score bounds.

3.4.1 Absence of External Data Sources

When there are no external data sources available, we prove a sufficient condition for being able to bound the counterfactual scores using only the biased data $\mathcal{D}_\Delta$. Note that we are not restricted to the binary attribute setting anymore. For calculating the necessity and sufficiency scores each, there are two cases of interest explored below. For each score, we take the maximum of the bound in two cases to determine the final result.

Proposition 3.2. *Given a PCM $\langle M', \Pr_\Delta(\mathbf{u})\rangle$ with a corresponding causal DAG $\mathcal{G}$ which represents the underlying biased data collection process, an algorithm $f : Dom(\mathbf{I}) \to Dom(O)$, and a set of attributes $\mathbf{X} \subseteq \mathbf{V} - \{O\}$ with two sets of attribute values $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$, if $\mathbf{K}$ consists of non-descendants of $\mathbf{I}$ in $\mathcal{G}$ and if $Pa(\mathbf{O})$, $\mathbf{O}$, and $\mathbf{X}$ are all disjoint from $Pa(C)$, then the explanation scores can be bounded as follows for any set of variables $\mathbf{K}$ such that $(\mathbf{O}, \mathbf{X} \perp\!\!\!\perp C \mid \mathbf{K})$:*

$$\textsc{Nec}_{\mathbf{x}}(\mathbf{k}) \leq \frac{1}{\Pr_\Delta(o, \mathbf{x} \mid \mathbf{k})} \cdot \max\{\textit{Nec. Case 1}, \textit{Nec. Case 2}\} \tag{9}$$

$$\textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \leq \frac{1}{\Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k})} \cdot \max\{\textit{Suff. Case 1}, \textit{Suff. Case 2}\} \tag{10}$$

$$\textsc{NeSuf}_{\mathbf{x}}(\mathbf{k}) \leq \Pr_\Delta(o, \mathbf{x} \mid \mathbf{k}) \cdot \textsc{Nec}_{\mathbf{x}}(\mathbf{k}) + \Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k}) \cdot \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \tag{11}$$

where, for necessity,

$$\begin{aligned}\textit{Nec. Case 1} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}\}} \Bigg(& \Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), Pa(\mathbf{O}), \mathbf{k}) \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o', \mathbf{x}' \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg) \\ & \cdot \Pr_\Delta(o \mid \mathbf{x}, Pa(\mathbf{O}), \mathbf{k}) \cdot \max_{\mathbf{u}} \Big\{ \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \Big\}\end{aligned}$$

and similarly, for the second case:

$$\begin{aligned}\textit{Nec. Case 2} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}\}} \Bigg(& \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o', \mathbf{x}' \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg) \\ & \cdot \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o \mid \mathbf{x}, Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \cdot \max_{\mathbf{u}} \Big\{ \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \Big\}\end{aligned}$$

and for sufficiency:

$$\begin{aligned}\textit{Suff. Case 1} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}'\}} \Bigg(& \Pr_\Delta(o \mid \texttt{do}(\mathbf{x}), Pa(\mathbf{O}), \mathbf{k}) \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o, \mathbf{x} \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg) \\ & \cdot \Pr_\Delta(o' \mid \mathbf{x}', Pa(\mathbf{O}), \mathbf{k}) \cdot \max_{\mathbf{u}} \Big\{ \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \Big\}\end{aligned}$$

and similarly, for the second case:

$$\textit{Suff. Case 2} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}\}} \Bigg(\max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o \mid \texttt{do}(\mathbf{x}), Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\}$$

$$- \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o, \mathbf{x} \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Big)$$
$$\cdot \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{ \Pr_\Delta(o' \mid \mathbf{x}', Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \} \cdot \max_{\mathbf{u}} \Big\{ \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \Big\}$$

Proof. We prove the bounds for (9). The proofs for (10) and (11) are similar.

$$\begin{aligned}
\text{LHS} &= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&\quad \texttt{(obtained from the consistency rule)} \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \frac{\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k})}{\Pr(\mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k})} \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \frac{\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x} \mid \text{Pa}(\mathbf{O}), \mathbf{k})}{\Pr(\mathbf{x} \mid \text{Pa}(\mathbf{O}), \mathbf{k})} \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \frac{\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \text{Pa}(\mathbf{O}), \mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k})}{\Pr(\mathbf{x} \mid \text{Pa}(\mathbf{O}), \mathbf{k})} \cdot \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \frac{\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \text{Pa}(\mathbf{O}), \mathbf{k}) - \Pr(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k})}{\Pr(\mathbf{x} \mid \text{Pa}(\mathbf{O}), \mathbf{k})} \cdot \Pr(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&\quad \texttt{(obtained from the consistency rule)} \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \frac{\Pr(o' \mid \texttt{do}(x'), \text{Pa}(\mathbf{O}), \mathbf{k}) - \Pr(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k})}{\Pr(\mathbf{x} \mid \text{Pa}(\mathbf{O}), \mathbf{k})} \cdot \Pr(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}), \mathbf{x} \mid \mathbf{k}) \\
&= \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \big(\Pr(o' \mid \texttt{do}(x'), \text{Pa}(\mathbf{O}), \mathbf{k}) - \Pr(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k}) \big) \cdot \Pr(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}) \cdot \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k})
\end{aligned}$$

Now, we can adopt the CRAB analysis in the case of no external information. Let $\mathbf{U} \subseteq \mathbf{X} \times \mathbf{O}$ where $\mathbf{U} \cap (\text{Pa}(\mathbf{O}) \cup \mathbf{K}) = \varnothing$, such that the following conditional independence is satisfied:

$$(\text{Pa}(\mathbf{O}) \perp\!\!\!\perp C \mid \mathbf{U}, \mathbf{K})$$

Here, we are also implicitly assuming that $\text{Pa}(C) \cap \text{Pa}(\mathbf{O}) = \varnothing$, because if variables in the parents of $\mathbf{O}$ directly influenced the selection variable, then by definition, the conditional independence assumption would not hold.

$$\begin{aligned}
\Pr\left(\text{Pa}(\mathbf{O}) \mid \mathbf{k}\right) &= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr(\mathbf{u} \mid \mathbf{k}) \\
&= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}, C = 1) \cdot \Pr(\mathbf{u} \mid \mathbf{k}) \\
&\leq \sum_{\mathbf{u} \in Dom(\mathbf{U})} \max_{\mathbf{u}} \{ \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}, C = 1) \} \cdot \Pr(\mathbf{u} \mid \mathbf{k}) \\
&= \max_{\mathbf{u} \in Dom(\mathbf{U})} \{ \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}, C = 1) \} \cdot \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr(\mathbf{u} \mid \mathbf{k}) \\
&= \max_{\mathbf{u} \in Dom(\mathbf{U})} \{ \Pr(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}, C = 1) \} \\
&= \max_{\mathbf{u} \in Dom(\mathbf{U})} \{ \Pr_\Delta(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \}
\end{aligned}$$

Substituting into the previous result yields:

$$\text{LHS} \leq \sum_{\text{Pa}(\mathbf{O})\setminus\{\mathbf{x}\}} \left(\underbrace{\Pr(o' \mid \texttt{do}(\mathbf{x}'), \text{Pa}(\mathbf{O}), \mathbf{k})}_{\text{Term I}} - \underbrace{\Pr(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k})}_{\text{Term II}} \right) \cdot \underbrace{\Pr(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k})}_{\text{Term III}} \cdot \max_{\mathbf{u} \in Dom(\mathbf{U})} \{\Pr_{\Delta}(\text{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u})\}$$

We still need to convert Terms I, II, and III, which are probabilities over the true population distribution, to the biased data distribution.

There are two cases of interest.

1. If $\mathbf{O}$ has no children, the conditional independence $\Pr(\mathbf{O} \perp\!\!\!\perp C \mid \text{Pa}(\mathbf{O}))$ holds trivially. We can directly change Terms I and III to the biased data distribution, and we will have to use the CRAB bound on Term II.
2. If $\mathbf{O}$ has children, we must use the CRAB bound on all of the terms.

Observe that since we already condition on $\text{Pa}(\mathbf{O})$ and $\mathbf{K}$, we do not have to condition on as many other variables in the support to achieve the conditional independence between $(\mathbf{O}, \mathbf{X})$ and the selection variable C.

Case 1. For Terms I and III, we can directly change the probabilities to be over the biased data distribution. We will have that:

$$\text{Term I} = \Pr_{\Delta}(o' \mid \texttt{do}(\mathbf{x}'), \text{Pa}(\mathbf{O}), \mathbf{k})$$

and

$$\text{Term III} = \Pr_{\Delta}(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k})$$

For Term II, we apply the CRAB analysis as follows. Define the disjoint subsets $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4 \subseteq \mathbf{U}$, such that all are disjoint from $\text{Pa}(\mathbf{O})$ and $\mathbf{K}$. Consider the graphical construction below that abides by the aforementioned conditions:

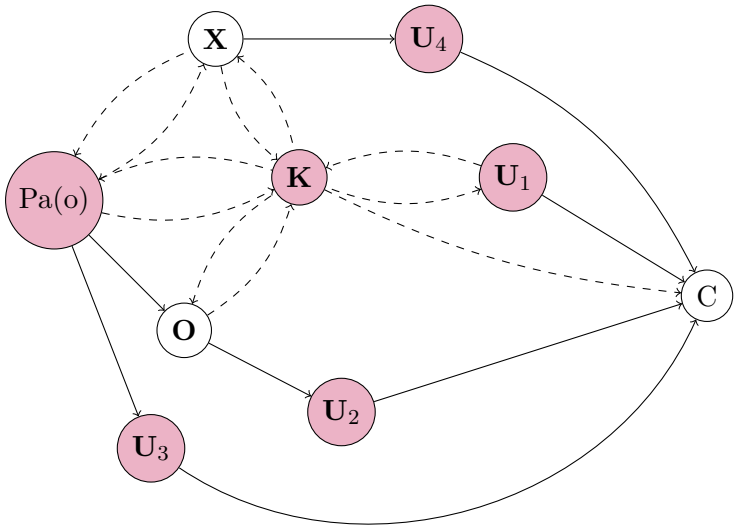


Figure 1: Bounding probabilities with only $\mathbf{O}$

In the above figure, shading of the nodes represents conditioning on those variables. A dashed arrow from one node to another indicates the possibility for a directed edge to exist between those two nodes. For clarity, we only include *some* of the possible dashed arrows possible, but including others (e.g., from $\mathbf{X}$ to $\mathbf{U_4}$) will not affect

our arguments or results. Note that any combination of these dashed arrows, along with all of the solid arrows, constitute a valid graph.

From the above model, observe that we will have to condition on $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_4$, as well as $\text{Pa}(\mathbf{O})$ and $\mathbf{K}$, in observe that in order to apply the CRAB bound as follows:

$$\text{Term II} \leq \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\}$$

Case 2. As in the previous case, we can apply the CRAB analysis to get the following bounds:

$$\begin{aligned}
\text{Term I} &\leq \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \\
\text{Term II} &\leq \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o', \mathbf{x}' \mid \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \\
\text{Term III} &\leq \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\}
\end{aligned}$$

However, in certain scenarios, it may be possible to further simplify Terms Term I and III. Consider the setting of Figure 1, as depicted below:

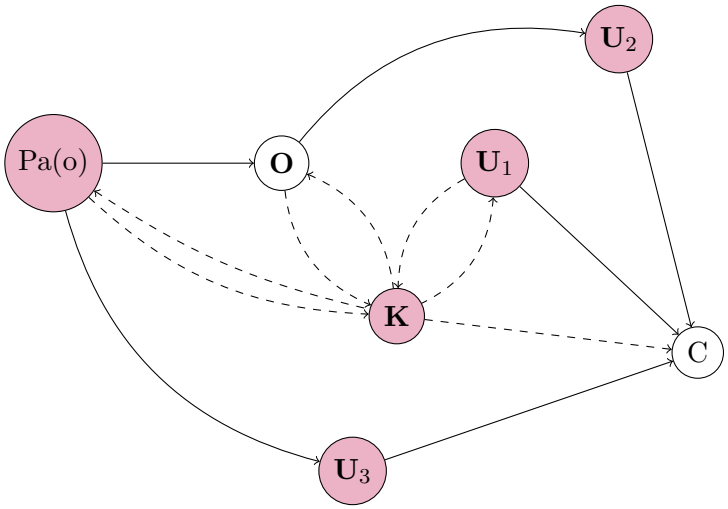


Figure 2: Bounding probabilities with only o

Note that Figure 1 is a generalization of Figure 2. In this scenario, we can form the conditional independence $(\mathbf{O} \perp\!\!\!\perp C \mid \mathbf{U}_1, \mathbf{U}_2, \text{Pa(o)})$. By applying the CRAB analysis, we have:

$$\text{Term I} \leq \max_{\mathbf{u}_1,\mathbf{u}_2} \{\Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2)\}$$

and

$$\text{Term III} \leq \max_{\mathbf{u}_1,\mathbf{u}_2} \{\Pr_\Delta(o \mid \mathbf{x}, \text{Pa}(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2)\}$$

Note that since we already condition on Pa(o) and $\mathbf{K}$, we do not have to condition on $\mathbf{U}_3$.

□

3.4.2 Presence of External Data Sources

Now we establish conditions under which scores can be tightly bounded from biased data when we have access to external data sources that reveal varying levels of information about the target population.

Proposition 3.3. *Given a PCM $\langle M', \Pr_\Delta(\mathbf{u})\rangle$ with a corresponding causal DAG $\mathcal{G}$ which represents the underlying biased data collection process, an algorithm $f : Dom(\mathbf{I}) \to Dom(O)$, and a set of attributes $\mathbf{X} \subseteq \mathbf{V} - \{O\}$ with two sets of attribute values $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$, if $\mathbf{K}$ consists of non-descendants of $\mathbf{I}$ in $\mathcal{G}$, if the auxiliary statistics $\Pr_\Omega(\mathbf{u} \mid \mathbf{k})$ for all $\mathbf{u} \in Dom(\mathbf{U})$ and $\mathbf{k} \in Dom(\mathbf{K})$ can be obtained using external data sources, and if $Pa(\mathbf{O})$, $\mathbf{O}$, and $\mathbf{X}$ are all disjoint from $Pa(C)$, then the explanation scores can be bounded as follows for any set of variables $\mathbf{K}$ such that $(\mathbf{O}, \mathbf{X} \perp\!\!\!\perp C \mid \mathbf{K})$:*

$$\textsc{Nec}_{\mathbf{x}}(\mathbf{k}) \leq \frac{1}{\Pr_\Delta(o, \mathbf{x} \mid \mathbf{k})} \cdot \max\{\textit{Nec. Case 1}, \textit{Nec. Case 2}\} \tag{12}$$

$$\textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \leq \frac{1}{\Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k})} \cdot \max\{\textit{Suff. Case 1}, \textit{Suff. Case 2}\} \tag{13}$$

$$\textsc{NeSuf}_{\mathbf{x}}(\mathbf{k}) \leq \Pr_\Delta(o, \mathbf{x} \mid \mathbf{k}) \cdot \textsc{Nec}_{\mathbf{x}}(\mathbf{k}) + \Pr_\Delta(o', \mathbf{x}' \mid \mathbf{k}) \cdot \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) \tag{14}$$

where, for necessity,

$$\begin{aligned}\textit{Nec. Case 1} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}\}} \Bigg(& \Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), Pa(\mathbf{O}), \mathbf{k}) \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o', \mathbf{x}' \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg) \\ & \cdot \Pr_\Delta(o \mid \mathbf{x}, Pa(\mathbf{O}), \mathbf{k}) \cdot \left(\sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr_\Omega(\mathbf{u} \mid \mathbf{k}) \right)\end{aligned}$$

and similarly, for the second case:

$$\begin{aligned}\textit{Nec. Case 2} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}\}} \Bigg(& \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o' \mid \texttt{do}(\mathbf{x}'), Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o', \mathbf{x}' \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg) \\ & \cdot \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \{\Pr_\Delta(o \mid \mathbf{x}, Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \cdot \left(\sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_\Delta(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr_\Omega(\mathbf{u} \mid \mathbf{k}) \right)\end{aligned}$$

and for sufficiency:

$$\begin{aligned}\textit{Suff. Case 1} \leq \sum_{Pa(\mathbf{O})\setminus\{\mathbf{x}'\}} \Bigg(& \Pr_\Delta(o \mid \texttt{do}(\mathbf{x}), Pa(\mathbf{O}), \mathbf{k}) \\ & - \max_{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4} \Big\{ \Pr_\Delta(o, \mathbf{x} \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \Big\} \Bigg)\end{aligned}$$

$$\cdot \Pr_{\Delta}(o' \mid \mathbf{x}', Pa(\mathbf{O}), \mathbf{k}) \cdot \left(\sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_{\Delta}(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr_{\Omega}(\mathbf{u} \mid \mathbf{k}) \right)$$

and similarly, for the second case:

$$\begin{aligned} \textit{Suff. Case 2} \leq \sum_{Pa(\mathbf{O}) \setminus \{\mathbf{x}\}} \Bigg(& \max_{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4} \{\Pr_{\Delta}(o \mid \texttt{do}(\mathbf{x}), Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \\ & - \max_{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4} \left\{ \Pr_{\Delta}(o, \mathbf{x} \mid Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4) \right\} \Bigg) \\ & \cdot \max_{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4} \{\Pr_{\Delta}(o' \mid \mathbf{x}', Pa(\mathbf{O}), \mathbf{k}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_4)\} \cdot \left(\sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_{\Delta}(Pa(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr_{\Omega}(\mathbf{u} \mid \mathbf{k}) \right) \end{aligned}$$

Proof. We prove the bounds for (12); (13) and (14) are proved similarly. The following result is obtained from the law of total expectation and the independence assumption $(\mathrm{Pa}(\mathbf{O}) \perp\!\!\!\perp C \mid \mathbf{U}, \mathbf{K})$, which requires that $\mathrm{Pa}(C) \cap \mathrm{Pa}(\mathbf{O}) = \varnothing$.

$$\begin{aligned} \Pr\left(\mathrm{Pa}(\mathbf{O}) \mid \mathbf{k}\right) &= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr(\mathrm{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr(\mathbf{u} \mid \mathbf{k}) \\ &= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr(\mathrm{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}, C = 1) \cdot \Pr(\mathbf{u} \mid \mathbf{k}) \\ &= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_{\Delta}(\mathrm{Pa}(\mathbf{O}) \mid \mathbf{k}, \mathbf{u}) \cdot \Pr_{\Omega}(\mathbf{u} \mid \mathbf{k}) \end{aligned}$$

The rest of the analysis follows precisely from the previous subsection (in the case of no external information). □

Note that while $\Pr_{\Delta}(\mathrm{Pa}(\mathbf{O}))$ can be estimated even from biased data, the external information must be used to compute $\Pr_{\Omega}(\mathbf{u} \mid \mathbf{k})$. The same independence assumption is used in borrowing from the previous analysis. By setting $\mathbf{U} = \mathrm{Pa}(C)$, i.e., the parents of the selection variable C in the biased data collection diagram $\mathcal{G}$, we can always find variables that satisfy this assumption. Therefore, these bounds are applicable even when the entire biased data collection diagram $\mathcal{G}$ is unknown and we have access to information only about $\mathrm{Pa}(C)$. Similar to the CRAB paper, we can select a minimal set of variables $\mathbf{U}$ (under different assumptions) that satisfies the independence condition and for which auxiliary information is available. This can be particularly useful when the external information does not cover the entire set of variables in $\mathrm{Pa}(C)$.

The previous results examine different ends of the spectrum in terms of the availability of external data. The former section requires no external data, while the current section so far requires sufficient external data for exact computation of a fairness query. In practice, one may have access to a level of external data that falls in between these two extremes. In such cases, it is important to note that the selection variable, C, may depend on a high-dimensional set of variables, and thus the set of variables $\mathbf{U}$ for which the conditions of the previous results hold could consist of a high-dimensional set of variables. This may make it infeasible to collect auxiliary information for computing all the statistics $\Pr_{\Omega}(\mathbf{u} \mid \mathbf{k})$ needed for tighter computation of counterfactual scores. Thus, in this case, we investigate the middle of the spectrum, where some auxiliary information about the target population is available but not enough for applying equations (12) through (14). We show that this limited about of auxiliary information can be used to compute a tighter upper bound for the scores than established in Section 3.4.2. Specifically, we consider similar assumptions as in the current section, but in situations where external data sources only have partial information about $\Pr_{\Omega}(\mathbf{u} \mid \mathbf{k})$.

4 Notes on Future Work

5 Appendix

5.1 Missing Proofs from Section 3.3

Proposition 5.1. *Explanation scores are related through the following inequality. For a binary X, the inequality becomes an equality.*

$$\textsc{NeSuf}_{\mathbf{x}}(\mathbf{k}) \leq \Pr(o, \mathbf{x} \mid \mathbf{k})\ \textsc{Nec}_{\mathbf{x}}(\mathbf{k}) + \Pr(o', \mathbf{x}' \mid \mathbf{k})\ \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) + 1 - \Pr(\mathbf{x} \mid \mathbf{k}) - \Pr(\mathbf{x}' \mid \mathbf{k}) \tag{15}$$

Proof. The inequality is obtained from the law of total probability, the consistency rule in (2), and applying the Fréchet bound, as shown in the following steps:

$$\begin{aligned}
\textsc{NeSuf}_{\mathbf{x}}(\mathbf{k}) &= \frac{\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}}, o'_{X\leftarrow x'}, \mathbf{k})}{\Pr(\mathbf{k})} \texttt{(from consistency (2))} \\
&= \frac{1}{\Pr(\mathbf{k})}\left(\sum_{x\in\{\mathbf{x},\mathbf{x}'\}} \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}}, o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{k}, x) + \sum_{x''\in Dom(X)-\{x,x'\}} \Pr(o_{X\leftarrow x}, o'_{X\leftarrow x'}, \mathbf{k}, \mathbf{x}'')\right) \\
&\qquad\qquad \texttt{(from law of total probability)} \\
&\leq \frac{1}{\Pr(\mathbf{k})}\left(\Pr(o_{X\leftarrow x}, o'_{\mathbf{X}\leftarrow\mathbf{x}}, \mathbf{k}, x) + \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}}, o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{k}, x') + \sum_{\mathbf{x}''\in Dom(X)-\{\mathbf{x},\mathbf{x}'\}} \Pr(\mathbf{k}, x'')\right) \\
&\qquad\qquad \texttt{(obtained by upper bounding the sum)} \\
&\leq \frac{1}{\Pr(\mathbf{k})}\left(\Pr(o, o'_{\mathbf{X}\leftarrow\mathbf{x}'}, \mathbf{k}, \mathbf{x}) + \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}}, o', \mathbf{k}, \mathbf{x}') + \Pr(\mathbf{k}) - \Pr(\mathbf{x}, \mathbf{k}) - \Pr(\mathbf{x}', \mathbf{k})\right) \\
&\leq \Pr(o, \mathbf{x} \mid \mathbf{k})\ \textsc{Nec}_{\mathbf{x}}(\mathbf{k}) + \Pr(o', \mathbf{x}' \mid \mathbf{k})\ \textsc{Suf}_{\mathbf{x}}(\mathbf{k}) + 1 - \Pr(\mathbf{x} \mid \mathbf{k}) - \Pr(\mathbf{x}' \mid \mathbf{k})
\end{aligned}$$

□

Therefore, for binary attributes, the necessary and sufficiency score can be seen as the weighted sum of necessary and sufficiency scores. Hence, if the causal effect of X on the algorithm's decision is non-zero, then so is the necessity and sufficiency score (for a binary X, it is implied from (15) that at least one of the sufficiency and necessity scores must also be non-zero).

References

[1] S. Galhotra and J. Y. Halpern. Intervention and conditioning in causal bayesian networks. *arXiv preprint arXiv:2405.14728*, 2024.

[2] S. Galhotra, R. Pradhan, and B. Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1234–1245. ACM, 2021.

[3] J. Pearl. *Causality*. Cambridge university press, 2009.

[4] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[5] J. Zhu, S. Galhotra, N. Sabri, and B. Salimi. Consistent range approximation for fair predictive modeling. *arXiv preprint*, 2022.