Causal Information Splitting: Engineering Proxy Features for Robustness to Distribution Shifts (Supplementary Material)

Bijan Mazaheri[1] **Atalanti Mastakouri**[2] **Dominik Janzing**[2] **Michaela Hardt**[2]

[1]California Institute of Technology, Pasadena, CA, USA
[2]Amazon Web Services, Tübingen, Germany

A ADDITIONAL SCM BACKGROUND

For any DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, we call $\boldsymbol{P} \subseteq \boldsymbol{E}$ a **path** if it connects A and B with no repeated vertices. The path is **directed** if it obeys the directions of the edges and **undirected** if it does not. Both directed and undirected paths in a causal DAG can result in dependencies between variables. To understand the conditions for dependence/independence (d-connection/d-separation) Pearl [1988], we will use the concepts of **active** and **inactive** paths, which are defined relative to a conditioning set Pearl [2009], Peters et al. [2017]. Intuitively, whether a path is active or not indicates whether it "carries dependence" between the variables.

For an empty conditioning set, a path between A to B is active if it is directed or if it is made up of two directed paths from a common cause along that path. In the same unconditioned setting, **inactive paths** are paths that contain a **collider**, i.e. a vertex for which the path has two inward pointing arrows.

When we are given a conditioning set $\boldsymbol{Z}$, conditional dependencies differ from unconditional ones. Active paths can be **blocked** (thus becoming inactive paths) if some vertex Z along the path between A to B is included in $\boldsymbol{C}$. Similarly, inactive paths with a collider C can become **unblocked** by including C or some descendant of the collider variable in the conditioning set $\boldsymbol{Z}$. If two variables A, B contain no active paths (they may contain inactive paths), then we say they are d-separated ($A \perp\!\!\!\perp_d B \mid \boldsymbol{Z}$). If two variables contain at least one active path for a conditioning set $\boldsymbol{Z}$, we say that they are d-connected.

Pearl [1988] uses structural causal models to justify the local Markov condition, which means that d-Separation always implies independence and allows DAG structures to be factorized. It is possible that two d-connected variables by chance exhibit some unexpected statistical independence. The assumption of faithfulness Spirtes et al. [2000] ensures that d-connectedness implies statistical dependence. This assumptions is popular in the causal discovery literature, but we will not need it for our theoretical results.

B INFORMATION THEORY PRELIMINARIES

Lemma 1 (Chain Rule, [Cover, 1999])**.** *For sets of variables $\boldsymbol{A}, \boldsymbol{B}$, and subset $\boldsymbol{B}' \subset \boldsymbol{B}$*

$$\mathcal{I}(\boldsymbol{A} : \boldsymbol{B}) = \mathcal{I}(\boldsymbol{A} : \boldsymbol{B}') + \mathcal{I}(\boldsymbol{A} : \boldsymbol{B} \setminus \boldsymbol{B'} \mid \boldsymbol{B}') \tag{1}$$

Definition 1 ([Cover, 1999])**.** For sets of variables $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, the **interaction information** is defined,

$$\mathcal{I}(\boldsymbol{A} : \boldsymbol{B} : \boldsymbol{C}) := \mathcal{I}(\boldsymbol{A} : \boldsymbol{B}) - \mathcal{I}(\boldsymbol{A} : \boldsymbol{B} \mid \boldsymbol{C}). \tag{2}$$

A key property of interaction information is that it is symmetric to permutations in its three inputs,

$$\mathcal{I}(\boldsymbol{A} : \boldsymbol{B} : \boldsymbol{C}) = \mathcal{H}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \mathcal{H}(\boldsymbol{A}) + \mathcal{H}(\boldsymbol{B}) + \mathcal{H}(\boldsymbol{C}) - \mathcal{H}(\boldsymbol{A}, \boldsymbol{B}) - \mathcal{H}(\boldsymbol{B}, \boldsymbol{C}) - \mathcal{H}(\boldsymbol{C}, \boldsymbol{A}). \tag{3}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Another key property is that interaction information can be either positive or negative, differing from mutual information which is non-negative. The following lemmas will describe two common situations in which we can expect positive and negative interaction information.

Lemma 2. *Given three sets of random variables $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ if $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$ then $\mathcal{I}(\boldsymbol{A} : \boldsymbol{B} : \boldsymbol{C}) \geq 0$.*

Graphically, Lemma 2 represents a situation where conditioning on $\boldsymbol{B}$ d-separates $\boldsymbol{A}$ and $\boldsymbol{C}$ (i.e. $\boldsymbol{B}$ is a <u>separating set</u> of $\boldsymbol{A}$ and $\boldsymbol{B}$). Hence, a sufficient condition to utilize Lemma 2 is $\boldsymbol{A} \mathrel{\circ\!\!\rightsquigarrow\!\!|} \boldsymbol{B} \mathrel{|\!\!\leftsquigarrow\!\!\circ} \boldsymbol{C}$, in which case the inequality becomes strict. The symmetry of interaction information means that it is not important which set of variables is the separating set. Conveniently $\mathcal{I}(A : B : C) \geq 0 \Rightarrow \mathcal{I}(A : B \mid C) \leq \mathcal{I}(A : B)$, meaning we can "drop" conditioned variables in our upper bounds.

The **data processing inequality** uses each "step" of an active path to upper bound the mutual information.

Lemma 3 (Data Processing Inequality (modified from Cover [1999]))**.** *If $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}, \boldsymbol{D}$ then*

$$\begin{aligned} \mathcal{I}(\boldsymbol{A} : \boldsymbol{C} \mid \boldsymbol{D}) &\leq \min(\mathcal{I}(\boldsymbol{A} : \boldsymbol{B} \mid \boldsymbol{D}), \mathcal{I}(\boldsymbol{B} : \boldsymbol{C} \mid \boldsymbol{D})) \\ &\leq \mathcal{H}(\boldsymbol{B} \mid \boldsymbol{D}). \end{aligned} \tag{4}$$

C REDUNDANCY BOUNDS WITHOUT FUNCTIONAL ASSUMPTIONS

Without the functional assumptions of our framework, we can only provide upper bounds on the context sensitivity.

To decompose the mutual information between sets of vertices $\boldsymbol{M}$ we will use the chain rule:

$$\mathcal{I}(Y : \boldsymbol{M} \mid \boldsymbol{X}) = \mathcal{I}(Y : M \mid \boldsymbol{X}) + \mathcal{I}(Y : \boldsymbol{M} \setminus \{M\} \mid \boldsymbol{X}, M) \tag{5}$$

Repeated application of the chain rule accumulates conditioned M variables for an arbitrary ordering. This conditioning represents the interactive nature of distribution shift mechanisms. We will now consider the context sensitivity of an arbitrary context variable conditioned on some arbitrary set of previously considered $\boldsymbol{M}' \subset \boldsymbol{M}$. We will be able to drop this conditioning on $\boldsymbol{M}'$ in our context sensitivity bounds.

Lemma 4. *For some $U_i \in \boldsymbol{U}$, and $\boldsymbol{M}' \subseteq \boldsymbol{M}$, if $M_i \mathrel{\circ\!\!\rightsquigarrow\!\!|} U_i \mathrel{|\!\!\leftsquigarrow\!\!\circ} Y$, then*

$$\mathcal{I}(M_i : Y \mid \boldsymbol{X}, \boldsymbol{M}') \leq \mathcal{H}(U_i \mid \boldsymbol{X}) = \mathcal{H}(U_i) - \mathcal{I}(U_i : \boldsymbol{X}).$$

Proof. Using the data processing inequality (Lemma 3),

$$\begin{aligned} &\mathcal{I}(M_i : Y \mid \boldsymbol{X}, \boldsymbol{M}') \\ &\qquad \leq \min(\mathcal{I}(U_i : M_i \mid \boldsymbol{X}, \boldsymbol{M}'), \mathcal{I}(U_i : Y \mid \boldsymbol{X}, \boldsymbol{M}')) \\ &\qquad \leq H(U_i \mid \boldsymbol{X}, \boldsymbol{M}') \\ &\qquad \leq H(U_i \mid \boldsymbol{X}) \end{aligned}$$

□

Lemma 4 gives an information-theoretic quantification of the notion of d-separation, showing that reducing context sensitivity involves selecting $\boldsymbol{X}$ with high $\boldsymbol{U}^{\text{GOOD}}$ redundancy. An invariant set is an extreme case of this rule, in which $\boldsymbol{X}$ has full redundancy with $\boldsymbol{U}^{\text{GOOD}}$.

Lemma 5. *For $U_i \in \boldsymbol{U}, \boldsymbol{X} \subseteq \boldsymbol{V}$ let $\boldsymbol{X}' = \boldsymbol{X} \cap \mathbf{CH}(U_i)$. If $U \in \boldsymbol{U}^{\text{BAD}}$ then*

$$\mathcal{I}(M_i : Y \mid \boldsymbol{X}, \boldsymbol{M}') \leq \mathcal{I}(U_i : \boldsymbol{X}' \mid Y) \leq \mathcal{I}(U_i : \boldsymbol{X}').$$

Lemma 5 quantitatively states that we should avoid redundancy with $\boldsymbol{U}^{\text{BAD}}$.

Proof. $M_i \perp\!\!\!\perp_d Y$ by the conditions of the lemma. Because $\boldsymbol{X} \setminus \boldsymbol{X}'$ has no vertices in $\mathbf{CH}(U_i)$ and there are no other descendants, $M_i \perp\!\!\!\perp_d Y \mid (\boldsymbol{X} \setminus \boldsymbol{X}')$.

$$\begin{aligned} \mathcal{I}(M_i : Y \mid \boldsymbol{X}, \boldsymbol{M}') &= -\mathcal{I}(M_i : Y : \boldsymbol{X}' \mid \boldsymbol{M}') \\ &= \mathcal{I}(M_i : \boldsymbol{X}' \mid Y, \boldsymbol{M}') \\ &\quad - \mathcal{I}(M_i : \boldsymbol{X} \mid \boldsymbol{M}') \\ &\leq \mathcal{I}(M_i : \boldsymbol{X}' \mid Y, \boldsymbol{M}'). \end{aligned} \tag{6}$$

$M_i \mathrel{\circ\!\!\rightsquigarrow\!|} U_i \mathrel{|\!\leftsquigarrow\!\!\circ} \boldsymbol{X}'$ because U_i is the only descendant of M_i. The DPI gives,

$$\mathcal{I}(M_U : \boldsymbol{X}' \mid Y, \boldsymbol{M}') \leq \mathcal{I}(U : \boldsymbol{X}' \mid Y, \boldsymbol{M}'). \tag{7}$$

We apply Lemma 2 because $\boldsymbol{X}' \mathrel{\circ\!\!\rightsquigarrow\!|} U_i \mathrel{|\!\leftsquigarrow\!\!\circ} \boldsymbol{M}'$

$$\mathcal{I}(M_U : \boldsymbol{X}' \mid Y, \boldsymbol{M}') \leq \mathcal{I}(U : \boldsymbol{X}' \mid Y). \tag{8}$$

We also have $\boldsymbol{X}' \mathrel{\circ\!\!\rightsquigarrow\!|} U_i \mathrel{|\!\leftsquigarrow\!\!\circ} Y$ from structural sparsity, so applying Lemma 2 gives the final loosest bound. □

D GUARANTEEING REDUNDANCY

Though we cannot ever measure the relevance of a proxy with unobserved $\boldsymbol{U}$, Lemma 6 shows that we can harness the graphical structure to obtain lower bounds.

Lemma 6 (Unobserved common-cause information)**.** *Given a causal DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, for any $U, V_i, V_j \in \boldsymbol{V}$ that satisfy $V_i \mathrel{\circ\!\!\rightsquigarrow\!|} U \mathrel{|\!\leftsquigarrow\!\!\circ} V_j$, then $\mathcal{I}(V_i, V_j : U) \geq \mathcal{I}(V_i : V_j)$.*

Proof. A visualization of this proof is given in Figure 1. Colors are added to the equations in the proof to match this figure. Begin with the definition of mutual information.

$$\mathcal{I}(V_i, V_j : U) = \mathcal{H}(V_i, V_j) - \mathcal{H}(V_i, V_j \mid U) \tag{9}$$

We can expand the joint entropy of both terms as follows,

$$\mathcal{H}(V_i, V_j) = \mathcal{H}(V_i \mid V_j) + \mathcal{H}(V_j \mid V_i) + \mathcal{I}(V_i : V_j) \tag{10}$$

$$\mathcal{H}(V_i, V_j \mid U) = \mathcal{H}(V_i \mid V_j, U) + \mathcal{H}(V_j \mid V_i, U) + \underbrace{\mathcal{I}(V_i : V_j \mid U)}_{=0 \text{ because } V_i \perp\!\!\!\perp_d^{\mathcal{G}} V_j | U} \tag{11}$$

Together, Equations 10 and 11 give:

$$\begin{aligned}\mathcal{I}(V_i, V_j : U) =& \mathcal{H}(V_i \mid V_j, U) + \mathcal{H}(V_j \mid V_i) + \mathcal{I}(V_i : V_j) - \mathcal{H}(V_i \mid U, V_j) - \mathcal{H}(V_j \mid U, V_i) \\ =& \mathcal{I}(V_i : U \mid V_j) + \mathcal{I}(V_j : U \mid V_i) + \mathcal{I}(V_i : V_j) \\ \geq& \mathcal{I}(V_i : V_j)\end{aligned}$$

□

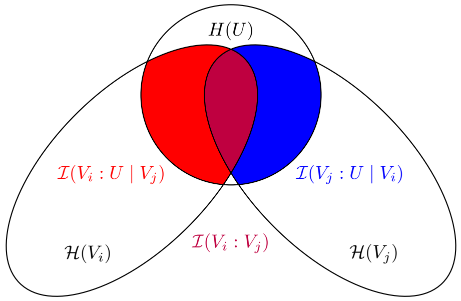

Figure 1: A visual proof for Lemma 6.

Hence, we can lower bound $F_{\text{ISO}(V_G)}(U_A)$'s hold on good information in $U_A^{(G)}$ using Lemma 6,

$$\mathcal{I}((V_i, F_{\text{ISO}(V_i)}(V_A)) : \mathbf{PA}(V_i)) \geq \mathcal{I}(V_A : V_i). \tag{12}$$

E DEFERRED PROOFS

E.1 PROOF OF LEMMA 1

Proof. We can rewrite the conditional mutual information making use of $U_1 \perp\!\!\!\perp U_2$ as follows.

$$\begin{aligned}\mathcal{I}(U_1 : U_2 \mid V) &= -\mathcal{I}(U_1 : U_2 : V) \\ &= -\mathcal{I}(U_1 : V) + \mathcal{I}(U_1 : V \mid U_2) \\ &= -\mathcal{I}(U_1 : V^{(U_1)}) + \mathcal{I}(U_1 : V^{(U_2)} \mid U_2) \\ &= -\mathcal{I}(U_1 : V^{(U_1)}) + \mathcal{I}(U_1 : V^{(U_1)}) = 0.\end{aligned}$$

□

E.2 PROOF OF LEMMA 2

Proof. $\mathcal{I}(U : \boldsymbol{X}) = \mathcal{H}(U) - \mathcal{H}(U \mid \boldsymbol{X})$. If at least one child of U is conditioned on and transmits (i.e. in $V \in \mathbf{CH}(U) \cap \boldsymbol{X}$ and $v_{\boldsymbol{x}} \neq \phi$), then $\mathcal{H}(U \mid \boldsymbol{x}) = 0$. Otherwise, $\mathcal{H}(U \mid \boldsymbol{X}) = \mathcal{H}(U)$ because all $X \in \boldsymbol{X} \setminus \mathbf{CH}(U)$ with active paths to U must go through colliders in $\mathbf{CH}(U)$ - all of which are not in $\boldsymbol{X}$ or not transmitting. □

E.3 PROOF OF LEMMA 3

Proof. $\mathcal{I}(M_i : Y \mid \boldsymbol{X}) = \sum_{\boldsymbol{X}} \mathcal{I}(M_i : Y \mid \boldsymbol{X})$ is zero unless (M_i, U_i) and (U_i, Y) both transmit. Furthermore, the DPI gives $\mathcal{I}(M_i : Y \mid \boldsymbol{X}) \leq \mathcal{H}(U_i \mid \boldsymbol{X})$, which is zero if any of the edges from $U_i \rightarrow X$ for $X \in \boldsymbol{X}$ transmit. □

E.4 PROOF OF LEMMA 4

Proof. $\mathcal{I}(M_i : Y \mid \boldsymbol{X}) = \sum_{\boldsymbol{X}} \mathcal{I}(M_i : Y \mid \boldsymbol{X})$ is zero unless both (M_i, U_i) and (U_i, Y) transmit and at least one of (U_i, X) transmits for $X \in \boldsymbol{X}$, in which case $\mathcal{I}(M_i : Y \mid \boldsymbol{x}) = \mathcal{I}(M_i : Y \mid U_i)$. □

E.5 PROOF OF LEMMA 5

Proof. ($\Rightarrow$) We will prove this with the contrapositive. If there is no shared parent between V_i and V_j, then all active paths must go through Y. However, because at least one of V_i, V_j is not connected to a cause, all paths between V_i and V_j cannot have a collider at Y (through two causes). This means conditioning on Y blocks the remaining paths.

($\Leftarrow$) If there is a shared parent U between V_i and V_j, then $V_i \leftarrow U \rightarrow V_j$ is an active path, d-connecting the two vertices. If V_i and V_j each have corresponding parents $U_i, U_j \in \mathbf{PA}(Y)$, then $V_i \leftarrow U_i \rightarrow Y \leftarrow U_j \rightarrow V_j$ is an active path conditioned on Y. □

E.6 PROOF OF LEMMA 6

Proof. The proof follows from Lemma 5. Adjacent edges in $\mathcal{G}_Y$ either indicate shared parents or that both vertices have a (potentially different) cause of Y as their parent.

If V_i and V_j share a parent, then $\mathbf{PA}(V_i) \subseteq \boldsymbol{U}^{\text{GOOD}}$ implies $\boldsymbol{U}^{\text{GOOD}} \cap \mathbf{PA}(V_j) \neq \emptyset$, so V_j has at least one "good" parent. The symmetric argument holds for $\mathbf{PA}(V_i) \subseteq \boldsymbol{U}^{\text{BAD}}$.

If V_i and V_j both have at least one causal parent, then we know $V_i, V_j \notin \boldsymbol{V}^{\text{BAD}}$. We know both vertices have at least one good U as a parent, so it trivially follows that both are either in $\boldsymbol{V}^{\text{GOOD}}$ or $\boldsymbol{V}^{\text{AMBIG}}$. □

E.7 PROOF OF THEOREM 1

Proof. The only potential conditioned collider is Y, which is not allowed to be separable by partial faithfulness. Hence, partial faithfulness guarantees that we can construct $\mathcal{G}_Y$ from conditional independence tests because d-connection implies dependence between V.

The requirements on $\boldsymbol{V}^*$ given by the theorem ensure that every $V \in \boldsymbol{V}$ has at least one label from an adjacency to a $V^* \in \boldsymbol{V}^*$ in $\mathcal{G}_Y$.

The algorithm adds "good" labels to all vertices with a known good parent and "bad" labels to all vertices with a known bad parent. Therefore, all "ambigious" vertices are correctly labeled.

We now only need to guarantee that that the "good" and "bad" vertices are not ambiguous. If V were ambiguous, it would be connected to a U of the opposite label (i.e. a "good" vertex would be connected to a bad U). Such a U would have at least one $V^* \in \boldsymbol{V}^* \cap \mathbf{CH}(U)$ which would be adjacent to V and have given V the label of U, a contradiction. □

E.8 PROOF OF LEMMA 7

Proof. $\mathcal{I}(M_i : Y \mid \boldsymbol{X}) = \sum_{\boldsymbol{x} \in \boldsymbol{X}} \Pr(\boldsymbol{x}) \mathcal{I}(M_i : Y \mid \boldsymbol{x})$. Now, we have

$$\mathcal{I}(M_i : Y \mid \boldsymbol{x}) = \mathcal{H}(M_i \mid \boldsymbol{x}) = \mathcal{H}(U_i \mid \boldsymbol{x})$$

if both (M_i, U_i) and (U_i, Y_i) edges transmit, which occurs with probability $\alpha_{M_i,U_i}\alpha_{U_i,Y}$. Pulling this coefficient outside of the sum gives $\mathcal{I}(M_i : Y \mid \boldsymbol{X}) = \alpha_{M_i,U_i}\alpha_{U_i,Y}\, \mathcal{H}(U_i \mid \boldsymbol{x})$.

□

E.9 PROOF OF LEMMA 8

Proof. $U_i \in \boldsymbol{U}^{\mathrm{BAD}}$ means $M_i \circ\!\!\!\rightsquigarrow U_i \leftsquigarrow\!\!\!\circ Y$, so $\mathcal{I}(M_i : Y) = 0$.

$$\begin{aligned} \mathcal{I}(M_i : Y \mid \boldsymbol{X}) &= -\mathcal{I}(M_i : Y : \boldsymbol{X}) \\ &\leq \mathcal{I}(M_i : \boldsymbol{X} \mid Y) \\ &\leq \mathcal{I}(U_i : \boldsymbol{X} \mid Y) = 0 \end{aligned} \tag{13}$$

The final inequality comes from the data processing inequality. □

E.10 PROOF OF LEMMA 9

Proof. Consider the function $F_C(V_A) = F_C(G, B)) = F_{\mathrm{ISO}(V_G)}(G)$. By definition,

$$\mathcal{I}(F_{\mathrm{ISO}(V_G)}(G) : V_G \mid Y) = \mathcal{I}(G : V_G \mid Y) \tag{14}$$

$$= \mathcal{I}(V_A : V_G \mid Y). \tag{15}$$

Hence, F_C is in the feasible set of the optimization function defining isolation functions. Furthermore, $F_C(V_A)$ is only a function of G and $G \perp\!\!\!\perp U_B \mid Y$, so we can also conclude that $F_C(V_A) \perp\!\!\!\perp U_B$. This means

$$\begin{aligned} \mathcal{H}(F_C(V_A) \mid Y) &= \mathcal{H}(F_C(V_A) \mid U_B, Y) \\ &\leq \mathcal{H}(F_{\mathrm{ISO}(V_G)}(V_A) \mid U_B, Y) \\ &\leq \mathcal{H}(F_{\mathrm{ISO}(V_G)}(V_A) \mid Y) - \mathcal{I}(F_{\mathrm{ISO}(V_G)}(V_A) : U_B \mid Y). \end{aligned}$$

Hence, if $\mathcal{I}(F_{\mathrm{ISO}(V_G)}(V_A) : U_B \mid Y) > 0$, then $\mathcal{H}(F_C(V_A) \mid Y) < \mathcal{H}(F_{\mathrm{ISO}(V_G)}(V_A) \mid Y)$, contradicting the minimality of $F_{\mathrm{ISO}(V_G)}(V_A)$. □

E.11 PROOF OF THEOREM 2

Proof. To shorten some equations, we will use

$$\boldsymbol{F}(V_A) := F_{\text{ISO}(V_G)}(V_A \mid Y).$$

We first show that Equation 8 is sufficient for an improvement in relevance. We can expand the relevance of $\boldsymbol{X}^+$ as follows

$$\begin{aligned}\mathcal{I}(Y : \boldsymbol{X}^+) =& \mathcal{I}(Y : \boldsymbol{F}(V_A)) + \mathcal{I}(Y : V_G \mid \boldsymbol{F}(V_A)) \\ &\geq \mathcal{I}(Y : V_G \mid \boldsymbol{F}(V_A)) \\ &\geq \mathcal{I}(Y : V_G) - \mathcal{I}(Y : V_G : \boldsymbol{F}(V_A)).\end{aligned} \tag{16}$$

So, for guaranteed improvement in relevance ($\mathcal{I}(Y : \boldsymbol{X}^+) > \mathcal{I}(Y : V_G)$), we need negative $\mathcal{I}(Y : V_G : \boldsymbol{F}(V_A)) < 0$. Expanding,

$$\mathcal{I}(Y : V_G : \boldsymbol{F}(V_A)) = \mathcal{I}(\boldsymbol{F}(V_A) : V_G) - \mathcal{I}(\boldsymbol{F}(V_A) : V_G \mid Y). \tag{17}$$

Thus, Equation 8 gives us the exact condition needed for negative interaction information, guaranteeing improvement.

We can show that the context sensitivity is no worse by separately considering the context sensitivity with $\mathbf{PA}(\boldsymbol{U}^{\text{BAD}})$ and $\mathbf{PA}(\boldsymbol{U}^{\text{GOOD}})$. We begin with $\boldsymbol{U}^{\text{BAD}}$. Applying Lemma 9,

$$\mathcal{I}((V_G, \boldsymbol{F}(V_A)) : \boldsymbol{U}^{\text{BAD}} \mid Y) = 0, \tag{18}$$

which satisfies the conditions for Lemma 8 to ensure us that $\mathcal{I}(\mathbf{PA}(\boldsymbol{U}^{\text{BAD}}) : Y \mid \boldsymbol{X}^+) = 0$.

Now, consider an arbitrary for $M_G = \mathbf{PA}(U_G) \in \mathbf{PA}(\boldsymbol{U}^{\text{GOOD}})$. Lemma 7 tells us that

$$\mathcal{I}(M_G : Y \mid \boldsymbol{X}^+) = \alpha_{M_G : U_G} \alpha_{U_G, Y} H(U_G \mid \widetilde{\mathbf{CH}}_{\boldsymbol{X}^+}(U_G)).$$

We then observe that $H(U_G \mid \widetilde{\mathbf{CH}}_{\boldsymbol{X}^+}(U_G)) \leq H(U_G \mid \boldsymbol{X})$ because entropy is submodular, which leads us to conclude,

$$\mathcal{I}(M_G : Y \mid \boldsymbol{X}^+) \leq \mathcal{I}(M_G : Y \mid \boldsymbol{X}).$$

This completes the proof. □

F EXPERIMENTS

F.1 SYNTHETIC EXPERIMENTAL SETUP

M_G and M_B are drawn from normal distributions with mean 0 and variable standard deviations. All other vertices (other than Y) are the average of their parents plus additional Gaussian noise $N(0, .2)$. $T_A \in \mathbb{R}^2$ is generated by applying a rotation matrix to $(T_A^{(G)}, T_A^B)^T$[1]. Y indicates whether its parents sum to a positive number with a 5% probability of flipping randomly.

F.2 F1 SCORES FOR REAL WORLD EXPERIMENT

We give the F1 scores for the experiment described in Section 7.2 in Table 1.

F.3 REAL WORLD EXPERIMENT IN-DOMAIN PERFORMANCE

Here we provide the results of the in-domain accuracy for the experiment described in Sec. 7.2. Recall, that we use US Census data and consider distributions shifts across time as suggested by Ding et al. [2021]. Table 2 shows the accuracy on 2019 data on a held-out dataset (separate from the training split). We repeated the experiment 10 times on different training/testing splits and report the mean and standard deviation of the accuracy for the largest states in the U.S. As expected, using all features has the most predictive power for in-domain tasks.

[1] Many rotations were tried in our experiments with identical results, so we display results from a 45 degree rotation.

Table 1: Comparison of out-of-domain (2021) performance on predicting high income via F1 scores.

State	All Features	Engineered Features	Limited Features
CA	**0.684**	**0.683**	0.676
FL	**0.459**	0.388	0.388
GA	0.541	**0.626**	**0.624**
IL	0.563	**0.630**	**0.628**
NY	**0.688**	**0.690**	0.662
NC	**0.475**	0.410	0.410
OH	0.519	**0.581**	**0.580**
PA	0.531	**0.608**	**0.606**
TX	0.554	**0.619**	**0.619**
avg	0.557	**0.582**	0.577

Table 2: Comparison of in-domain (2019) performance on predicting high income via Accuracies.

State	All Features	Engineered Features	Limited Features
CA	**0.713** ± 0.0010	**0.710** ± 0.0012	0.691 ± 0.0011
FL	**0.700** ± 0.0014	0.693 ± 0.0020	0.694 ± 0.0017
GA	**0.708** ± 0.0025	**0.708** ± 0.0036	**0.707** ± 0.0036
IL	**0.689** ± 0.0023	**0.690** ± 0.0039	**0.685** ± 0.0021
NY	**0.705** ± 0.0024	0.698 ± 0.0022	0.687 ± 0.0076
NC	**0.713** ± 0.0020	0.703 ± 0.0049	0.700 ± 0.0028
OH	**0.717** ± 0.0029	**0.716** ± 0.0042	**0.712** ± 0.0033
PA	**0.702** ± 0.0028	**0.701** ± 0.0027	0.695 ± 0.0026
TX	**0.708** ± 0.0019	**0.705** ± 0.0025	**0.706** ± 0.0022
avg	**0.706**	0.703	0.697

References

Thomas M Cover. Elements of information theory. John Wiley & Sons, 1999.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems, 34, 2021.

Judea Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann, 1988.

Judea Pearl. Causality. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. Causation, prediction, and search. MIT press, 2000.