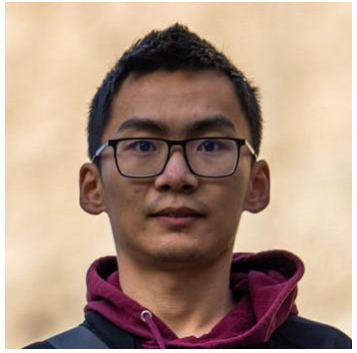
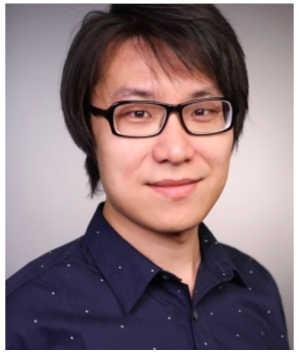
Procedural Fairness Through Decoupling Objectionable Data Generating Components

Zeyu Tang
zeyutang@cmu.edu



Jialu Wang
faldict@ucsc.edu



Yang Liu
yangliu@ucsc.edu

Peter Spirtes
ps7z@andrew.cmu.edu

Kun Zhang
kunz1@cmu.edu

Our Contribution in a Nutshell

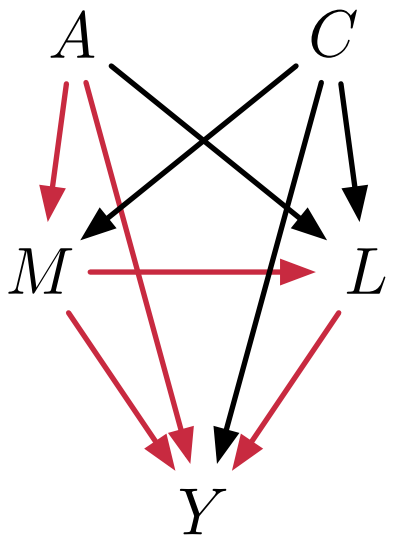
1. We reveal and address the often-overlooked issue of ***disguised procedural unfairness***

2. To decouple objectionable data generating components, we propose ***value instantiation rule***

3. We configure ***reference points*** to further satisfy requirements of procedural fairness

A Motivating Example - Model Setup

Causal graph

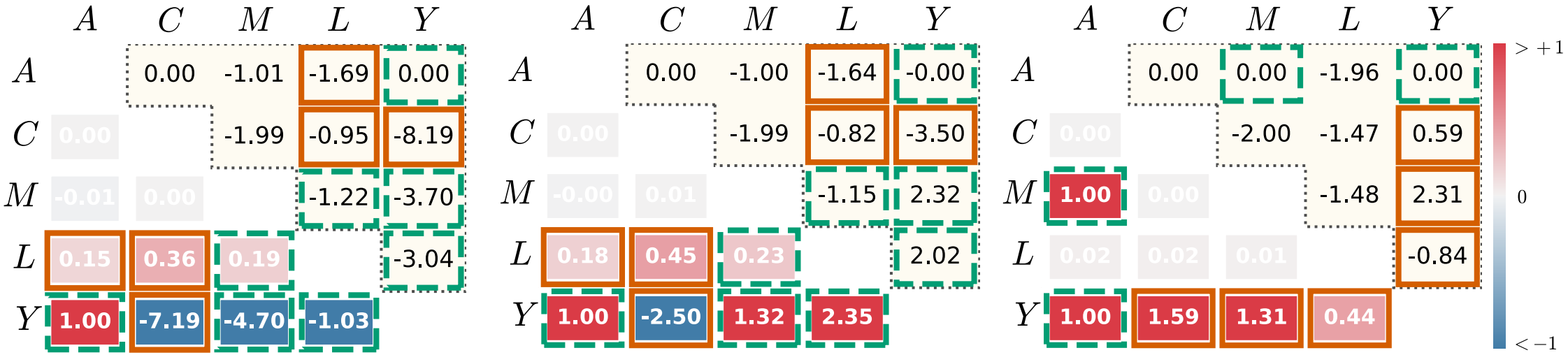


Functional causal model

$$
\begin{aligned}
A &\sim \text{Bernoulli}(p_A), \\
C &= \epsilon_C, \\
M &= \theta_A^M A + \theta_C^M C + \theta^M + \epsilon_M, \\
L &= \theta_A^L A + \theta_C^L C + \theta_M^L M + \theta^L + \epsilon_L, \\
Y &= \theta_A^Y A + \theta_C^Y C + \theta_M^Y M + \theta_L^Y L + \theta^Y + \epsilon_Y.
\end{aligned}
$$

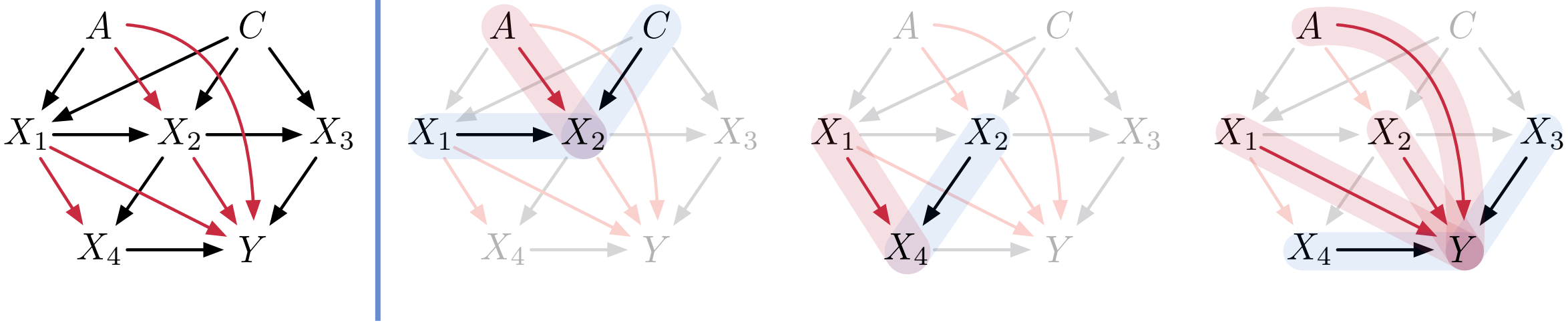
A Motivating Example - Illustration of Issue

- Previous proposals
  - causal fairness notions focus on causal effect(s) between A and Y
  - unintentional side-effect on neutral components



Our Framework - Value Instantiation Rule (1)

- Instead of constraints over model parameters,
  we find **appropriate input value** for **local causal modules**



Our Framework - Value Instantiation Rule (2)

Algorithm 1: The Value Instantiation Rule for Local Causal Modules

Input : The $d_{\text{in}}(V;\mathcal{G})$-ary function $h_V\left(\text{Parents}(V);\hat{\theta}_V\right)$ modeling the causal mechanism between the node V and its direct parents, where $d_{\text{in}}(V;\mathcal{G})$ is the the number of direct parents (in-degree) of V in the graph $\mathcal{G}$. The configuration function $\text{ReferencePoint}(\cdot)$, which maps a directed edge corresponding to an objectionable component $\rho \in \mathcal{E}_{\text{Obj}}$ to a reference point (Definition 4.1) with the domain of value of the tail node of the edge.

Output : The derivation of the predicted outcome $\widehat{V}$ in the local causal module.

If *there is additional assumption on the functional form* $\widetilde{h}_V(\cdot)$ *and/or parameters* $\widetilde{\theta}_V$ **Then**
  $\hat{\theta}_V \leftarrow \widetilde{\theta}_V, h_V \leftarrow \widetilde{h}_V$; `// direct correction of the causal mechanism`
Else
  ForEach *parent node* W_j *in* $\text{Parents}(V) = (W_1, W_2, \ldots, W_{d_{\text{in}}(V;\mathcal{G})})$ **Do**
    If *the edge* $\rho_j = (W_j, V) \in \mathcal{E}_{\text{Obj}}$, *i.e.,* $W_j \to V$ *is an objectionable component* **Then**
      w_j gets the value $\text{ReferencePoint}(\rho_j)$, because $W_j = \text{Tail}(\rho_j)$;
    Else If *there is at least one ancestor nodes of* W_j *was set to a reference point* **Then**
      w_j gets the value that W_j would have taken as a downstream of its ancestor nodes, to which reference points, if any, have been assigned;
    Else
      w_j gets the value of variable W_j for the record in the data set;
$\widehat{v} \leftarrow h_V(w_1, w_2, \ldots, w_{d_{\text{in}}(V;\mathcal{G})};\hat{\theta}_V)$.

Value instantiation option:
Reference point configuration

Value instantiation option:
Downstream of reference point(s)

Value instantiation option:
Original value in data set

Our Framework - Value Instantiation Rule (3)

Algorithm 1: The Value Instantiation Rule for Local Causal Modules

Input : The $d_{\text{in}}(V;\mathcal{G})$-ary function $h_V\left(\text{Parents}(V);\hat{\theta}_V\right)$ modeling the causal mechanism between the node V and its direct parents, where $d_{\text{in}}(V;\mathcal{G})$ is the the number of direct parents (in-degree) of V in the graph $\mathcal{G}$. The configuration function $\text{ReferencePoint}(\cdot)$, which maps a directed edge corresponding to an objectionable component $\rho \in \mathcal{E}_{\text{Obj}}$ to a reference point (Definition 4.1) with the domain of value of the tail node of the edge.

Output : The derivation of the predicted outcome $\widehat{V}$ in the local causal module.

If *there is additional assumption on the functional form* $\widetilde{h}_V(\cdot)$ *and/or parameters* $\widetilde{\theta}_V$ **Then**
  $\hat{\theta}_V \leftarrow \widetilde{\theta}_V, h_V \leftarrow \widetilde{h}_V$; `// direct correction of the causal mechanism`
Else
  ForEach *parent node* W_j *in* $\text{Parents}(V) = (W_1, W_2, \ldots, W_{d_{\text{in}}(V;\mathcal{G})})$ **Do**
    If *the edge* $\rho_j = (W_j, V) \in \mathcal{E}_{\text{Obj}}$, *i.e.,* $W_j \to V$ *is an objectionable component* **Then**
      w_j gets the value $\text{ReferencePoint}(\rho_j)$, because $W_j = \text{Tail}(\rho_j)$;
    Else If *there is at least one ancestor nodes of* W_j *was set to a reference point* **Then**
      w_j gets the value that W_j would have taken as a downstream of its ancestor nodes, to which reference points, if any, have been assigned;
    Else
      w_j gets the value of variable W_j for the record in the data set;
$\widehat{v} \leftarrow h_V(w_1, w_2, \ldots, w_{d_{\text{in}}(V;\mathcal{G})};\hat{\theta}_V)$.

Value instantiation option:
Reference point configuration

Value instantiation option:
Downstream of reference point(s)

Value instantiation option:
Original value in data set

Our Framework - Value Instantiation Rule (4)

Algorithm 1: The Value Instantiation Rule for Local Causal Modules

Input : The $d_{\text{in}}(V;\mathcal{G})$-ary function $h_V\left(\text{Parents}(V);\hat{\theta}_V\right)$ modeling the causal mechanism between the node V and its direct parents, where $d_{\text{in}}(V;\mathcal{G})$ is the the number of direct parents (in-degree) of V in the graph $\mathcal{G}$. The configuration function ReferencePoint$(\cdot)$, which maps a directed edge corresponding to an objectionable component $\rho \in \mathcal{E}_{\text{Obj}}$ to a reference point (Definition 4.1) with the domain of value of the tail node of the edge.

Output : The derivation of the predicted outcome $\widehat{V}$ in the local causal module.

If *there is additional assumption on the functional form* $\widetilde{h}_V(\cdot)$ *and/or parameters* $\widetilde{\theta}_V$ **Then**
  $\hat{\theta}_V \leftarrow \widetilde{\theta}_V, h_V \leftarrow \widetilde{h}_V$; `// direct correction of the causal mechanism`
Else
  ForEach *parent node* W_j *in* $\text{Parents}(V) = (W_1, W_2, \ldots, W_{d_{\text{in}}(V;\mathcal{G})})$ **Do**
    If *the edge* $\rho_j = (W_j, V) \in \mathcal{E}_{\text{Obj}}$, *i.e.,* $W_j \to V$ *is an objectionable component* **Then**
      w_j gets the value ReferencePoint(ρ_j), because $W_j = \text{Tail}(\rho_j)$;
    Else If *there is at least one ancestor nodes of* W_j *was set to a reference point* **Then**
      w_j gets the value that W_j would have taken as a downstream of its ancestor nodes, to which reference points, if any, have been assigned;
    Else
      w_j gets the value of variable W_j for the record in the data set;
$\widehat{v} \leftarrow h_V(w_1, w_2, \ldots, w_{d_{\text{in}}(V;\mathcal{G})};\hat{\theta}_V)$.

Value instantiation option:
Reference point configuration

Value instantiation option:
Downstream of reference point(s)

Value instantiation option:
Original value in data set

Our Framework - Value Instantiation Rule (5)

Algorithm 1: The Value Instantiation Rule for Local Causal Modules

Input : The $d_{\text{in}}(V;\mathcal{G})$-ary function $h_V\left(\text{Parents}(V);\hat{\theta}_V\right)$ modeling the causal mechanism between the node V and its direct parents, where $d_{\text{in}}(V;\mathcal{G})$ is the the number of direct parents (in-degree) of V in the graph $\mathcal{G}$. The configuration function $\text{ReferencePoint}(\cdot)$, which maps a directed edge corresponding to an objectionable component $\rho \in \mathcal{E}_{\text{Obj}}$ to a reference point (Definition 4.1) with the domain of value of the tail node of the edge.

Output : The derivation of the predicted outcome $\widehat{V}$ in the local causal module.

If *there is additional assumption on the functional form* $\widetilde{h}_V(\cdot)$ *and/or parameters* $\widetilde{\theta}_V$ **Then**
  $\hat{\theta}_V \leftarrow \widetilde{\theta}_V, h_V \leftarrow \widetilde{h}_V$; `// direct correction of the causal mechanism`
Else
  ForEach *parent node* W_j *in* $\text{Parents}(V) = (W_1, W_2, \ldots, W_{d_{\text{in}}(V;\mathcal{G})})$ **Do**
    If *the edge* $\rho_j = (W_j, V) \in \mathcal{E}_{\text{Obj}}$, *i.e.,* $W_j \to V$ *is an objectionable component* **Then**
      w_j gets the value $\text{ReferencePoint}(\rho_j)$, because $W_j = \text{Tail}(\rho_j)$;
    Else If *there is at least one ancestor nodes of* W_j *was set to a reference point* **Then**
      w_j gets the value that W_j would have taken as a downstream of its ancestor nodes, to which reference points, if any, have been assigned;
    Else
      w_j gets the value of variable W_j for the record in the data set;
$\widehat{v} \leftarrow h_V(w_1, w_2, \ldots, w_{d_{\text{in}}(V;\mathcal{G})}; \hat{\theta}_V)$.

Value instantiation option:
Reference point configuration

Value instantiation option:
Downstream of reference point(s)

Value instantiation option:
Original value in data set

Our Framework - Overall Pipeline

- The overall pipeline utilizes **reference point configurations** that are to the greatest benefit of the least advantaged individuals

Algorithm 2: Aggregating Local Causal Modules while Decoupling Objectionable Components

Input : The data set $\mathcal{D}$, the hypothesis class $\mathcal{H}$ and the parameter space Θ, the causal graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, the list of index $\mathcal{I}$ for all nodes $\mathbf{V}$. The set of edges $\mathcal{E}_{\text{Obj}}$ where each edge corresponds to an objectionable component. The $\text{ReferencePoint}(\cdot)$ configuration.

Output : The derivation of the predicted outcome $\widehat{Y}$ that *decouples* objectionable components from the data generating process, and *only* makes use of neutral components.

Sort the list of index $\mathcal{I}$ such that parent nodes, if any, appear before the node itself;

ForEach *node index* $i \in \mathcal{I}$ **Do** `// learn model parameters`

  If *the number of direct parents of node* V_i*, i.e., the in-degree,* $d_{\text{in}}(V_i; \mathcal{G}) > 0$ **Then**

    Fit model parameters in the local causal module between V_i and its direct parent nodes $\text{Parents}(V_i)$, without any fairness constraint:

    $h_{V_i}, \hat{\theta}_{V_i} \leftarrow \underset{\theta \in \Theta, h \in \mathcal{H}}{\text{argmin}} \mathcal{L}_{V_i}\big(h(\text{Parents}(V_i); \theta), V_i; \mathcal{D}\big)$, $\mathcal{L}_{V_i}$ is the loss function for V_i;

According to the sorted list of node index $\mathcal{I}$, apply the value instantiation rule (Algorithm 1) to each local causal module in sequence, and then derive prediction $\widehat{Y}$ according to Equation (3).

Summary

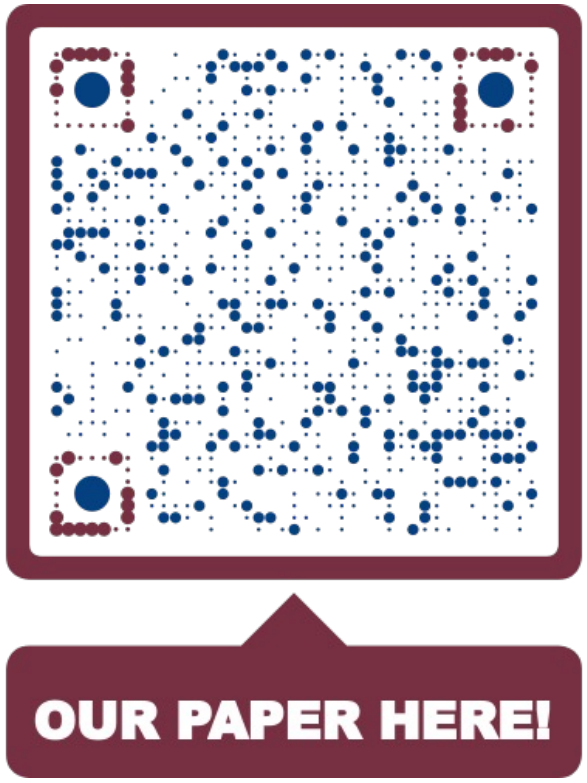
1. We reveal and address the often-overlooked issue of ***disguised procedural unfairness***

2. To decouple objectionable data generating components, we propose ***value instantiation rule***

3. We configure ***reference points*** to further satisfy requirements of procedural fairness

Thank you!

Our Paper



Our Code Repository