Algorithmic Recourse: from Counterfactual Explanations to Interventions

By Kamiri, Schölkopf, and Valera

Presentation by
Catherine Huang, Christina Xiao, Kelly Zhang

Algorithmic Recourse

- Increasingly algorithms are used to make consequential decisions for individuals

- Recourse: "*Systematic process of reversing unfavorable decisions made by algorithms and bureaucracies*"
    - Promoting **agency** and **trust**

Paper's Main Contributions

1. **Insufficiencies of previous problem formulations → Motivates causal approach to recourse (Kelly)**

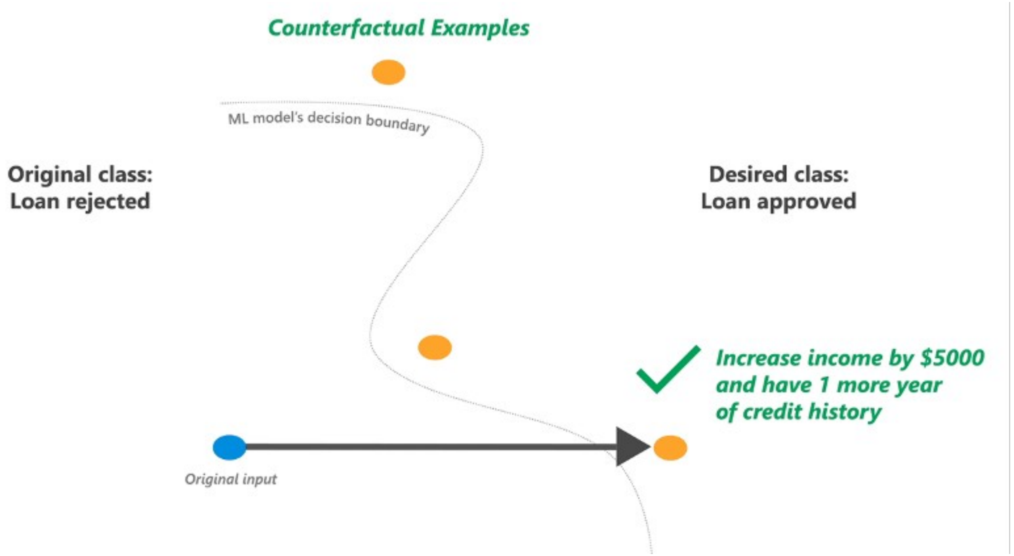
2. **Structural Causal Model approach to recourse (Catherine)**

3. **Discuss examples motivated by real-world problems and future directions (Christina)**

Nearest Counterfactual Explanations

- For a person with features $\mathbf{x}^F$ who was denied a loan, find the "nearest neighbor" $\mathbf{x}$ who was granted a loan

- Difference between $\mathbf{x}^F$ and $\mathbf{x}$ is an "explanation" for the loan denial for $\mathbf{x}^F$



Nearest Counterfactual Explanation (CFE)

$$\boldsymbol{x}^{*\mathsf{CFE}} \in \underset{\boldsymbol{x}}{argmin} \quad \mathrm{dist}(\boldsymbol{x}, \boldsymbol{x}^{\mathsf{F}}) \quad \text{s.t.} \quad h(\boldsymbol{x}) \neq h(\boldsymbol{x}^{\mathsf{F}}), \boldsymbol{x} \in \mathcal{P},$$

- **Finding nearest counterfactual explanation as an optimization problem**
  - $\mathbf{x}^{*\mathbf{CFE}}$ is the Counter Factual Explanation
  - Function **h** is the classifier

- **Distance metrics**
  - Lp norm;   L1 norm divided by median absolute deviation; etc

Why are counterfactual **explanations** not ideal for **recommendations** for recourse?

Counterfactual Explanations provide **understanding** but do not necessarily lead to optimal **action recommendations**

Why are counterfactual **explanations** not ideal for **recommendations** for recourse?

1. Don’t account for person’s difficulty or “cost” of changing dimensions of $\mathbf{x}^F$ (addressed by Ustun et al.)

2. Don’t account for downstream “causal” impact of taking actions (addressed by this work)

Accounting for the “Cost” of Actions (Ustun et al.)

$$\boldsymbol{\delta}^* \in \underset{\boldsymbol{\delta}}{\mathit{argmin}} \ \ \mathrm{cost}(\boldsymbol{\delta}; \boldsymbol{x}^{\mathsf{F}}) \quad \text{s.t. } h(\boldsymbol{x}^{\mathsf{CFE}}) \neq h(\boldsymbol{x}^{\mathsf{F}}),$$
$$\boldsymbol{x}^{\mathsf{CFE}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta},$$
$$\boldsymbol{x}^{\mathsf{CFE}} \in \mathcal{P}, \ \boldsymbol{\delta} \in \mathcal{F},$$

- δ^* restricted to the set of “feasible changes”
- Consider linear impact of changes δ
- Non-trivial to choose costs that reflect people's’ true objective functions

Insufficiencies of Ustun et al. Formulation

$$\boldsymbol{\delta}^* \in \underset{\boldsymbol{\delta}}{argmin} \ \ \mathrm{cost}(\boldsymbol{\delta}; \boldsymbol{x}^{\mathsf{F}}) \quad \text{s.t. } h(\boldsymbol{x}^{\mathsf{CFE}}) \neq h(\boldsymbol{x}^{\mathsf{F}}),$$

$$\boldsymbol{x}^{\mathsf{CFE}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta},$$

$$\boldsymbol{x}^{\mathsf{CFE}} \in \mathcal{P}, \ \ \boldsymbol{\delta} \in \mathcal{F},$$

1. Marginal cost of changing a feature is constant
   - **Example:** Cost of going from salary \$0 to \$100k equals cost to go from salary \$800k to \$900k

Insufficiencies of Ustun et al. Formulation

2. Doesn’t consider downstream “causal” impact of actions

    - **Example:** Loan decision for individual changed if “salary” reaches 100k (+33%) or “bank balance” reaches 30k (+20%).
      20% change seems easier.

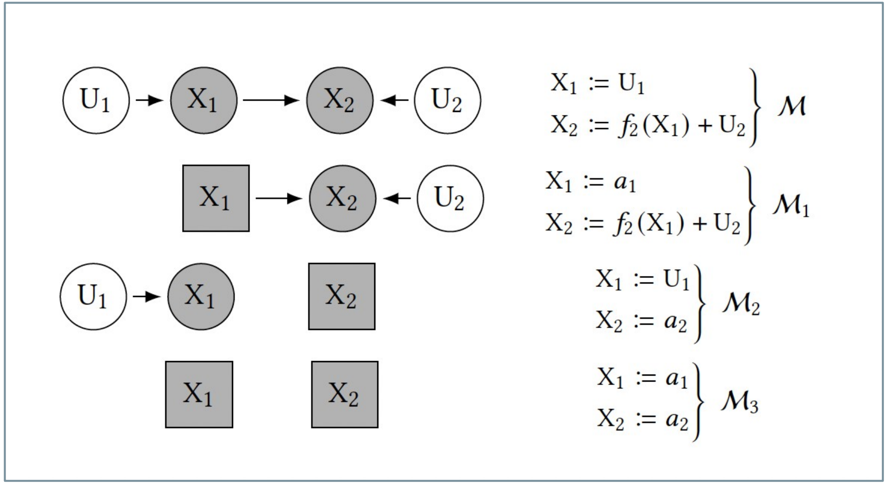
    - However, best **action recommendation** is to increase salary by 14% when 30% of salary automatically saved to bank.

Actions as Interventions: Structured Causal Model (SCM)

Setup

- M ($M \in \Pi$) = $\langle F, X, \mathrm{U} \rangle$: SCM
- X: endogenous (observed) variables
- U: exogenous (unobserved) variables
- F: U $\rightarrow$ X, structural equations
- A ($\Pi \rightarrow \Pi$): structural interventions, i.e. transformations between SCMs
    - of the form A := do($\{X_i := a_i\}_{i \in I}$)

Structural Interventions



- M: true world model
- observation: $M_1 \neq M_2 \neq M_3$

Actions as Interventions: Structural Counterfactuals

What will individual x^F 's feature vector be after the individual performs action set A in world M?

Assumptions: (1) no hidden confounders (true SCM), (2) full access to invertible F

Idea: once we know F, X (endogenous variables) can be uniquely determined given U (exogenous variables)

Compute $F^{-1}(x^F)$

Takeaway: we can compute any structural counterfactual query for individual x^F:

$$x^{SCF} = F_A(F^{-1}(x^F)).$$

Limitations of CFE-Based Recourse: Formalism

Setup: x^F (individual features), δ^* action recommendation (Ustun et al. solution), I (set of indices of acted-upon observed variables: $I = \{i \mid \delta^*_i \neq 0\}$)

Definition (CFE-Based Actions): a set of structural interventions $A^{CFE} := do(\{X_i := x_i^F + \delta^*\}_{i \in I})$

Proposition: $A^{CFE} \xrightarrow{\text{gives rise to}} x^{SCF} = x^{*CFE} := x^F + \delta^*$ (i.e. recourse is guaranteed) if and only if I's descendants = {}.

Corollary: if the true world M is independent—if all the observed features are root-nodes of G—then CFE-based actions always guarantee recourse.

Algorithmic Recourse via Minimal Interventions: Setup

Formulation

$$
\begin{aligned}
\mathrm{A}^* \in \underset{\mathrm{A}}{argmin} \quad & \mathrm{cost}(\mathrm{A}; \boldsymbol{x}^{\mathsf{F}}) \\
\text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\
& \boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathrm{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) \\
& \boldsymbol{x}^{\mathsf{SCF}} \in \mathcal{P}, \quad \mathrm{A} \in \mathcal{F},
\end{aligned}
$$

Remarks

- ~~finding minimal shift of features~~ → finding minimal cost action set that yields favorable label
- $A^* \in \mathfrak{F}$ = set of feasible actions with minimally costly recourse
- $\mathrm{cost}(\square; x^F): \mathfrak{F} \times X \to \mathbb{R}_+$
- x^{*SCF}: $F_{A^*}(F^{-1}(x^F))$: resulting counterfactual
- $x^{*SCF} \neq x^{*CFE}$ (from Ustun et al.)!

Algorithmic Recourse via Minimal Interventions: Formalism

(from the last slide)

$$
\begin{aligned}
\mathbf{A}^* \in \underset{\mathbf{A}}{argmin} \quad & \text{cost}(\mathbf{A}; \boldsymbol{x}^{\mathsf{F}}) \\
\text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\
& \boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) \\
& \boldsymbol{x}^{\mathsf{SCF}} \in \mathcal{P}, \quad \mathbf{A} \in \mathcal{F},
\end{aligned}
$$

Proposition:

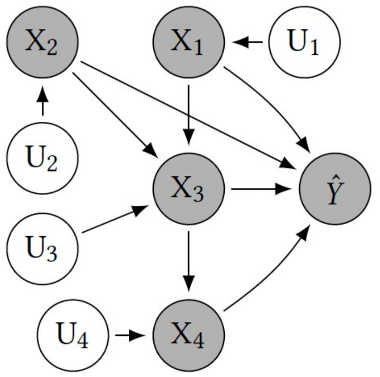
- A^{CFE} : Counter Factual Explanation-based action
- A* : Minimal Intervention Solution (blue box above)

$$\textbf{cost}(\mathbf{A}^*; x^{\mathsf{F}}) \leq \textbf{cost}(\mathbf{A}^{\mathsf{CFE}}; x^{\mathsf{F}})$$

Algorithmic Recourse via Minimal Interventions: MINT

- Recourse through Minimal Interventions (MINT) idea:
    - Required: that we can compute structural counterfactual of an individual in the world given *any* feasible action
    - Focus on the case where the SCM is an additive noise model
    - ⇒ Abduction-action-prediction technique (Pearl et al.) to compute x^{SCF}: $F_{\mathrm{A}}(F^{-1}(x^{\mathrm{F}}))$

Abduction-Action-Prediction to obtain x^{SCF}



$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= U_2 \\ X_3 &:= f_3(X_1, X_2) + U_3 \\ X_4 &:= f_4(X_3) + U_4 \end{aligned}\right\} \mathcal{M}$$

$$\hat{Y} = h\left(\{X_i\}_{i=1}^4\right)$$

$\{U_i\}^4_{i=1}$: mutually independent, exogenous

$\{f_i\}^4_{i=1}$: structural equations

$x = [x_1^{\mathrm{F}}, x_2^{\mathrm{F}}, x_3^{\mathrm{F}}, x_4^{\mathrm{F}}]^{\mathrm{T}}$: observed factual features

(1) Abduction: compute exogenous variables

$$\begin{aligned} u_1 &= x_1^{\mathrm{F}}, \\ u_2 &= x_2^{\mathrm{F}}, \\ u_3 &= x_3^{\mathrm{F}} - f_3(x_1^{\mathrm{F}}, x_2^{\mathrm{F}}), \\ u_4 &= x_4^{\mathrm{F}} - f_4(x_3^{\mathrm{F}}). \end{aligned}$$

(2) Action: modify SCM with interventions

$$\begin{aligned} X_1 &:= [1 \in I] \cdot a_1 + [1 \notin I] \cdot U_1, \\ X_2 &:= [2 \in I] \cdot a_2 + [2 \notin I] \cdot U_2, \\ X_3 &:= [3 \in I] \cdot a_3 + [3 \notin I] \cdot \left(f_3(X_1, X_2) + U_3\right), \\ X_4 &:= [4 \in I] \cdot a_4 + [4 \notin I] \cdot \left(f_4(X_3) + U_4\right), \end{aligned}$$

(3) Prediction: recursively compute endogenous variables based on (1) and (2)

$$\begin{aligned} x_1^{\mathrm{SCF}} &:= [1 \in I] \cdot a_1 + [1 \notin I] \cdot (u_1), \\ x_2^{\mathrm{SCF}} &:= [2 \in I] \cdot a_2 + [2 \notin I] \cdot (u_2), \\ x_3^{\mathrm{SCF}} &:= [3 \in I] \cdot a_3 + [3 \notin I] \cdot \left(f_3(x_1^{\mathrm{SCF}}, x_2^{\mathrm{SCF}}) + u_3\right), \\ x_4^{\mathrm{SCF}} &:= [4 \in I] \cdot a_4 + [4 \notin I] \cdot \left(f_4(x_3^{\mathrm{SCF}}) + u_4\right). \end{aligned}$$

General Formulation and Solving the Optimization Problem

General Formulation

$$
\begin{aligned}
\mathbf{A}^* \in \underset{\mathbf{A}}{\mathit{argmin}} \quad & \mathrm{cost}(\mathbf{A}; \boldsymbol{x}^{\mathsf{F}}) \\
\text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\
& \boxed{\begin{aligned} x_i^{\mathsf{SCF}} = {} & [i \in I] \cdot (x_i^{\mathsf{F}} + \delta_i) \\ & + [i \notin I] \cdot \left(x_i^{\mathsf{F}} + f_i(\mathrm{pa}_i^{\mathsf{SCF}}) - f_i(\mathrm{pa}_i^{\mathsf{F}})\right). \end{aligned}} \\
& \boldsymbol{x}^{\mathsf{SCF}} \in \mathcal{P}, \quad \mathbf{A} \in \mathcal{F},
\end{aligned}
$$

Remarks

- $x_i^{\mathsf{F}} + \delta_i$: intervention
- $f_i(\mathbf{pa}_i^{\mathsf{F}})$: factual values of x_i's parents
- $f_i(\mathbf{pa}_i^{\mathsf{SCF}})$: counterfactual values of x_i's parents
- new closed-form expression for $F_{A^*}(F^{-1}(x^{\mathsf{F}})) \Rightarrow$ use optimization methods

Experimental Setup

Synthetic setting:

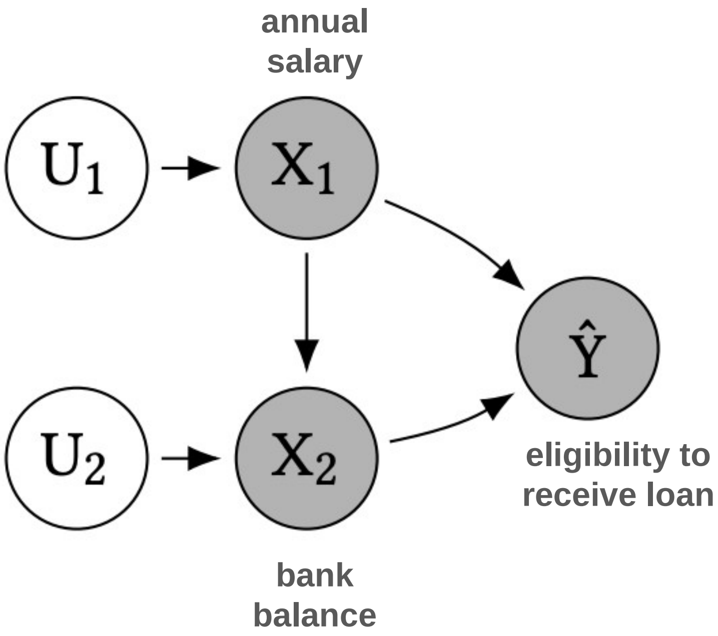
Generate data following causal generative process

Real-world setting:

Use existing German credit dataset to learn structural causal model equations, by fitting a linear regression

Cost for both is ℓ_1 norm over normalized feature change

Experimental Setup: Synthetic Setting



$U_1 \sim \$10000 \cdot \text{Poisson}(10),$

$U_2 \sim \$2500 \cdot N(0,1)$

$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= f_2(X_1) + U_2 \end{aligned}\right\} \mathcal{M}$$

$X_2 := 3/10 \cdot X_1 + U_2$

$$\hat{Y} = h(X_1, X_2)$$

$h = \text{sgn}(X_1 + 5 \cdot X_2 - \$225000)$

Casual Generative Process

Experimental Results: Synthetic Setting

[annual salary, bank balance]

$\mathbf{x}^F$		$[\$75000, \$25000]^T$
Karimi et al.'s formulation	**A***	$do(X_1 := x_1^F + \$10000)$
	$\mathbf{x}^{*SCF}$	$[\$85000, \$28000]^T$
Prior formulation	δ*	$[\$0, +\$5000]^T$
	$\mathbf{x}^{*CFE}$	$[\$75000, \$30000]^T$

Experimental Results: Synthetic Setting

[annual salary, bank balance]

$\mathbf{x}^F$		[\$75000, \$25000]^T
Karimi et al.'s formulation	$\mathbf{x}^{*SCF}$	[\$85000, \$28000]^T
Prior formulation	$\mathbf{x}^{*CFE}$	[\$75000, \$30000]^T

$\mathbf{x}^{*SCF}$ further dist from $\mathbf{x}^F$ than $\mathbf{x}^{*CFE}$

BUT

$\text{cost}(\delta^*; \mathbf{x}^F) \approx 2\ \text{cost}(\mathbf{A}^*; \mathbf{x}^F)$

Experimental Setup

Synthetic setting:

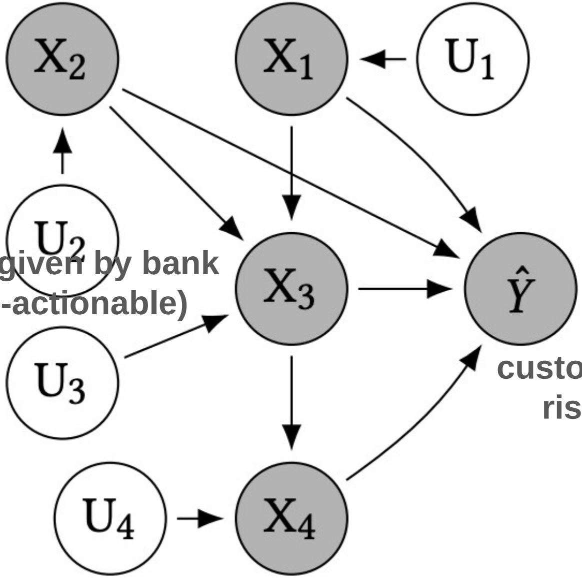
Generate data following causal generative process

Real-world setting:

Use existing German credit dataset to learn structural causal model equations, by fitting a linear regression

Cost for both is ℓ_1 norm over normalized feature change

Experimental Setup: Real-World Setting



$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= U_2 \\ X_3 &:= f_3(X_1, X_2) + U_3 \\ X_4 &:= f_4(X_3) + U_4 \end{aligned}\right\} \mathcal{M}$$

$$\hat{Y} = h\left(\{X_i\}_{i=1}^{4}\right)$$

h can be logistic regression or decision tree

Structural Causal Model

Experimental Results: Real-World Setting

[gender, age, credit given, credit repayment duration]

$\mathbf{x}^F$		$[\text{Male}, 32, \$1938, 24]^T$
Karimi et al.'s formulation	**A***	$do(\{X_2 := x_2^F + 1, X_3 := x_3^F - \$800\})$
	$\mathbf{x}^{*SCF}$	$[\text{Male}, 33, \$1138, 22]^T$
Prior formulation	δ*	$[\text{N/A}, +6, \$0, 0]^T$
	$\mathbf{x}^{*CFE}$	$[\text{Male}, 38, \$1938, 24]^T$

Experimental Results: Real-World Setting

[gender, age, credit given, credit repayment duration]

x^F		$[\text{Male}, 32, \$1938, 24]^T$
Karimi et al.'s formulation	**x^{*SCF}**	$[\text{Male}, 33, \$1138, 22]^T$
Prior formulation	**x^{*CFE}**	$[\text{Male}, 38, \$1938, 24]^T$

42% decrease in cost using Karimi et al.'s formulation

Averaged over 50 test individuals, 39 ± 24% and 65 ± 8% decrease in cost, for *h* as logistic regression and decision tree, respectively

Future Work: Extended Kinds of Interventions

Future Work: Extended Kinds of Interventions

Forms

- **Structural/hard** (actions in Karimi et al.)**:** unconditionally sever all edges incident on intervened node
- **Additive/soft:** do not sever incident edges

$$x_i^{\mathsf{SCF}} = [i \in I] \cdot \delta_i + \left(x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}})\right).$$

Future Work: Extended Kinds of Interventions

Forms

- **Structural/hard** (actions in Karimi et al.)**:** unconditionally sever all edges incident on intervened node
- **Additive/soft:** do not sever incident edges

$$x_i^{\mathsf{SCF}} = [i \in I] \cdot \delta_i + \left(x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}})\right).$$

Feasibility

Can encode as constraints to amend to $\mathbf{A} \in F$

- **Immutable:** closed under ancestral relationships; $[i \notin I] = 1$ recursively necessitates fulfillment of $[j \notin I] = 1$ for all $j \in \mathrm{pa}_i$
- **Mutable but non-actionable:** $[i \notin I] = 1$ is sufficient
- **Actionable and mutable:** contingent on (a) pre-intervention value of variable (b) pre-intervention value of other variables (c) post-intervention value of variable (d) post-intervention value of other variables

Future Work: Extended Kinds of Interventions

Forms

- **Structural/hard** (actions in Karimi et al.)**:** unconditionally sever all edges incident on intervened node
- **Additive/soft:** do not sever incident edges

$$x_i^{\mathsf{SCF}} = [i \in I] \cdot \delta_i + \left(x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}})\right).$$

Scopes

- Karimi et al. assumes action = intervention on endogenous variable
- **Fat-hand/non-atomic:** confounded/correlated interventions

Feasibility

Can encode as constraints to amend to $\mathbf{A} \in F$

- **Immutable:** closed under ancestral relationships; $[i \notin I] = 1$ recursively necessitates fulfillment of $[j \notin I] = 1$ for all $j \in \mathrm{pa}_i$
- **Mutable but non-actionable:** $[i \notin I] = 1$ is sufficient
- **Actionable and mutable:** contingent on (a) pre-intervention value of variable (b) pre-intervention value of other variables (c) post-intervention value of variable (d) post-intervention value of other variables

Future Work: Current Limitations

Reliance on true causal model of the world

Future Work: Current Limitations

True for any approach suggesting actions to be performed in the real world?

Reliance on true causal model of the world

Future Work: Current Limitations

Reliance on true causal model of the world

True for any approach suggesting actions to be performed in the real world?

Study potential inefficiencies from partial/imperfect causal model

Concluding Thoughts

- Overall an interesting work bridging counterfactual explanations (previous week’s papers) and algorithmic recourse (next week’s papers), clarifying their differences
- Tradeoff between generalizability of setting and hardness of optimization problem
  - Ustun et al. pursue algorithmic recourse in a specific linear setting
  - Karimi et al. pursue algorithmic recourse in general settings, require true causal model of world
  - Open question: where do we stand on this tradeoff? Consider the origins of counterfactual explanations in Wachter et al.