On the Adversarial Robustness of Causal Algorithmic Recourse

Ricardo Dominguez-Olmedo[1,2]
Amir-Hossein Karimi[1,3]
Bernhard Schölkopf[1,3]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany; [2]University of Tübingen, Tübingen, Germany; [3]ETH Zürich, Zürich, Switzerland;

Summary

- Algorithmic recourse aims to offer actionable recommendations for individuals to overcome unfavorable decisions made by ML classifiers. To warrant trust, recourse should be robust to reasonably small changes in the circumstances of the individual seeking recourse [1].
- Methods generating minimally costly recourse fail to be robust: small changes to the individual may invalidate the recourse action.
- We present methods to generate robust recourse recommendations in both the linear and the non-linear case.
- We derive bounds on the extra cost incurred by individuals seeking robust recourse, and we motivate a model regularizer for training classifiers such that the additional cost of robust recourse is reduced.

Notation & background (causality + recourse)

Causality We assume that the features $\mathbf{X} = \{X_1, \dots, X_d\}$ of individuals $x \in \mathcal{X}$ are governed by a known *structural causal model* [2] (SCM) $\mathcal{M} = (\mathbf{S}, \mathbb{P}_{\mathbf{U}})$:

- $\mathbf{S} = \{X_i := f_i(\mathsf{PA}_i, U_i)\}_{i=1}^d$ are the *structural equations* assigning each X_i as a function f_i of its direct causes (causal parents) $\mathsf{PA}_i \subseteq \mathbf{X} \setminus X_i$ and noise U_i
- $\mathbf{U} = \{U_i\}_{i=1}^d$ is a set of *unobserved* (exogenous/latent) *noise variables*
- $\mathbb{P}_{\mathbf{U}} = \mathbb{P}_{U_1} \times \dots \times \mathbb{P}_{U_d}$ is a factorising *noise distribution* ($\rightarrow$ causal sufficiency)

Counterfactuals allow to reason about what would have happened under certain hypothetical interventions all else being equal. We denote the mapping from factual to counterfactual as $\mathbb{CF}(x, do(\mathbf{X}_\mathcal{I} = \theta))$ or $\mathbb{CF}(x, \Delta)$, for

Hard interventions $do(\mathbf{X}_\mathcal{I} = \theta)$	*Additive interventions* Δ
Fix the values of a subset of features $\mathbf{X}_\mathcal{I}$	Perturb the features while preserving all causal relationships
$\mathbf{S}_{\mathcal{I}_i}^{do(\mathbf{X}_\mathcal{I}=\theta)} = \mathbf{X}_{\mathcal{I}_i} := \theta_i$	$\mathbf{S}^\Delta = \left\{X_i := f_i\left(\mathbf{X}_{\mathsf{pa}(i)}, \mathbf{U}_i\right) + \Delta_i\right\}_{i=1}^n$

Recourse [3]

- Recourse action: modelled as hard interventions $a = do(\mathbf{X}_\mathcal{I} = x_\mathcal{I} + \theta)$
- Classifier: assigns favourable/unfavourable outcomes $h : \mathcal{X} \rightarrow \{0, 1\}$
- Cost function: effort required to adopt recourse a $c(x, a) \in \mathbb{R}$
- Feasibility set: actions available to individual x $\mathcal{F}(x)$
- An action a is a valid recourse action if... $h(\mathbb{CF}(x, a)) = 1$

Minimum-cost recourse is fragile

Let a^* be the minimum-cost recourse action for the individual x. Under the following mild conditions

i) $c(x, do(X_\mathcal{I} = x_\mathcal{I} + \theta)$ is strictly convex in θ with minimum $\theta = 0$

ii) $do(X_\mathcal{I} = x_\mathcal{I} + \theta)) \in \mathcal{F}(x) \implies do(X_\mathcal{I} = x_\mathcal{I} + t\theta)) \in \mathcal{F}(x) \;\; \forall\, 0 < t < 1$

iii) The structural equations of the SCM $\mathcal{M}$ are continuous.

the counterfactual $\mathbb{CF}(x, a^*)$ must be precisely at the decision boundary. Thus, minimum-cost recourse can be invalidated by arbitrarily small changes to the features of the individual seeking recourse (see figure on top middle).

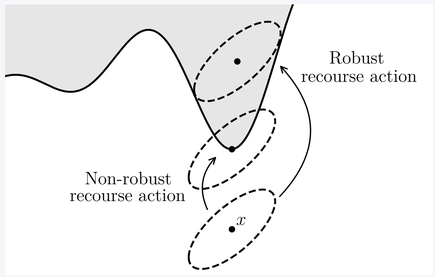
The adversarially robust recourse problem

The standard recourse problem seeks the minimum-cost recourse action a such that the corresponding counterfactual is favourably classified [4]

$$\underset{a = do(X_\mathcal{I} = x_\mathcal{I} + \theta)}{\arg\min} \; c(x, a) \quad \text{s.t.} \quad a \in \mathcal{F}(x) \;\wedge\; h\left(\mathbb{CF}(x, a)\right) = 1 \tag{1}$$

We guard against uncertainty by defining an uncertainty set $B(x)$ of individuals "similar" to x, and requiring robust recourse actions to be valid for all individuals in the uncertainty set $x' \in B(x)$

$$\underset{a = do(X_\mathcal{I} = x_\mathcal{I} + \theta)}{\arg\min} \; \underset{x' \in B(x)}{\max} \; c(x', a) \quad \text{s.t.} \quad a \in \mathcal{F}(x') \;\wedge\; h\left(\mathbb{CF}(x', a)\right) = 1 \tag{2}$$

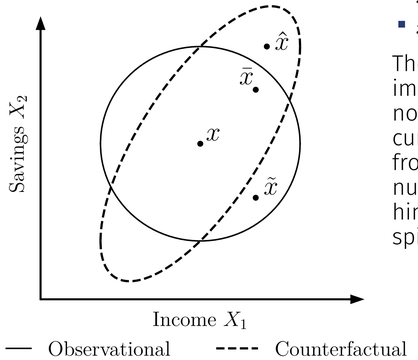


Counterfactual similarity

We define the uncertainty set $B(x)$ by adopting a causal view on neighbourhoods of similar individuals:

Observational neighbourhood	Counterfactual neighbourhood
Features of similar individuals may differ by ϵ-small perturbations δ	Perturbations are modelled as additive interventions on x
$B(x) = \{x + \delta \mid \|\|\delta\|\| \leq \epsilon\}$	$B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\|\Delta\|\| \leq \epsilon\}$

We argue that neighbourhoods of counterfactually similar individuals can be more informative, as the causal relationships between features are explicitly considered. For the SCM $X_1 = U_1$, $X_2 = X_1 + U_2$ with similarity neighbourhoods



- $\bar{x}$ has higher income, higher savings
- $\tilde{x}$ has higher income, lower savings

The counterfactual neighbourhood implies $d(x, \bar{x}) < d(x, \tilde{x})$. Since $\tilde{x}$ is not well explained by the SCM, its circumstances may substantially differ from those of x (e.g. has a much larger number of individuals dependent on him/her, resulting in lower savings despite its higher income).

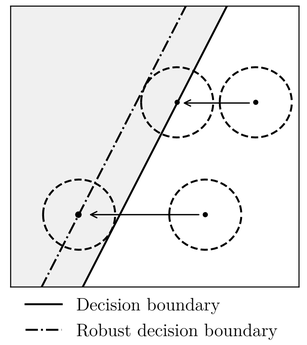
Generating robust recourse

The linear case For a linear classifier $h(x) = \langle w, x \rangle \geq b$ and linear SCM, seeking adversarially robust recourse is equivalent to seeking standard recourse for a modified classifier $h'(x) = \langle w, x \rangle \geq b'$ with shifted decision boundary $b' = b + ||J_{\mathbb{S}^\mathcal{I}}^T w||^*$, where $J_{\mathbb{S}^\mathcal{I}}$ denotes the Jacobian of the mapping resulting from hard-intervening on the features $\mathbf{X}_\mathcal{I}$. See figure below.

The non-linear case In general, the adversarially robust recourse problem can be solved by considering the unconstrained penalty objective

$$\mathcal{L}(x, a, \lambda) = c(x, a) + \lambda\, \ell\left(h\left(\mathbb{CF}\left(x, a\right)\right), 1\right) \tag{3}$$

A suitable algorithm is provided below. The inner maximization over the uncertainty set (Line 5) can be readily solved with projected gradient ascent.



Require: $\lambda > 0$, $\gamma > 1$
1: $\theta \leftarrow 0$
2: **while** $N \leq N_{\max}$ **do**
3: **while** not converged **do**
4: $a \leftarrow do(X_\mathcal{I} = x_\mathcal{I} + \theta)$
5: $x^* \leftarrow \arg\max_{x' \in B(x)} \mathcal{L}(x', a, \lambda)$
6: **if** $h\left(\mathbb{CF}\left(x^*, a\right)\right) = 1$ **then**
7: **return** θ
8: $\theta \leftarrow \mathrm{Proj}_{\mathcal{F}(x)}\left(\theta - \alpha \nabla_\theta \mathcal{L}(x^*, a, \lambda)\right)$
9: $\lambda \leftarrow \gamma\lambda$
10: **return** θ

A bound on the additional cost of robust recourse

For a linear classifier $h(x) = \langle w, x \rangle \geq b$ and an SCM with linear structural equations, if $a = do(\mathbf{X}_\mathcal{I} = x_\mathcal{I} + \theta)$ is a recourse action, we derive the following upper bound on the additional cost incurred to robustify action a

$$\frac{c(x, a') - c(x, a)}{c(x, a)} \leq \beta\epsilon, \quad \beta = \frac{||J_{\mathbb{S}^\mathcal{I}}^T w||^*}{\langle J_{\mathbb{S}^\mathcal{I}}^T w, \theta \rangle}, \; a' \text{ is robust recourse} \tag{4}$$

The weights w of the classifier determine the upper bound β on the additional cost of robust recourse. We propose to regularize the classifier in order to reduce the upper bound β, by penalizing the dual norm $||w_\mathcal{U}||^*$ of the weights $w_\mathcal{U}$ corresponding to unactionable features. This is theoretically well-motivated, as detailed in the paper.

For an experimental evaluation of the methods proposed please refer to our preprint in arXiv, which is an extension of this work.

[1] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020.
[2] Judea Pearl. *Causality*. Cambridge university press, 2009.
[3] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.
[4] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.