On the Adversarial Robustness of Causal Algorithmic Recourse

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

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

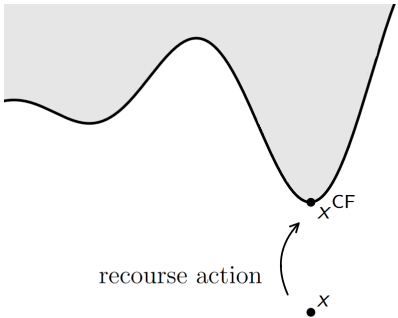
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Algorithmic recourse

Provide *actionable recommendations* to overcome unfavorable classification outcomes.

In a loan application setting:

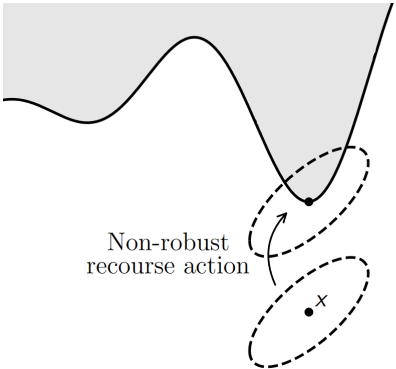
"If your monthly salary were to increase by \$1000, your loan would be approved."



The fragility of recourse

Theorem 1 *Minimum-cost recourse is fragile to arbitrarily small changes to the features of the individual x seeking recourse.*

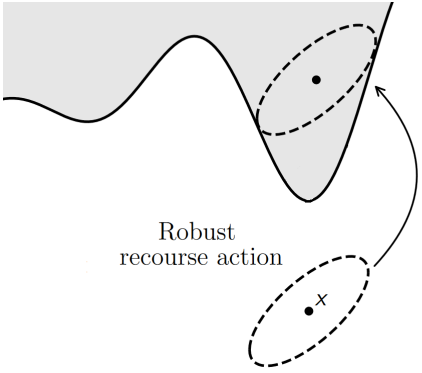
- e.g., a recommendation may no longer be valid if age changes by $\pm$ a month
- ... trustworthy recommendations?



Adversarially robust recourse

Iff results in favorable classification outcomes for all plausible individuals in some *uncertainty set* $B(x)$

$$h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \quad \forall x' \in B(x)$$



The adversarially robust algorithmic recourse problem

$$\begin{aligned} \underset{a=do(\mathbf{X}_{\mathcal{I}}=x_{\mathcal{I}}+\theta)}{\arg\min} \quad & c(x,a) && \text{Search for the least effortful action...} \\ \text{s.t.} \quad & a \in \mathcal{F}(x) && \text{that is actionable} \\ & h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \;\; \forall x' \in B(x) && \text{and leads to a favourable decision for all plausible individuals in the } \textit{uncertainty set } B(x) \end{aligned}$$

- We model a as causal interventions on the features x ... (Karimi et al., 2021)
- ... in order to reason about counterfactuals in a principled manner (Pearl, 2009)

Generating adversarially robust recourse

- **Linear** $h(x) = \langle w, x \rangle \geq b$, generate w.r.t. a larger acceptance threshold $b' \geq b$
- **Differentiable** h, solve optimization under *adversarial perturbations* to x

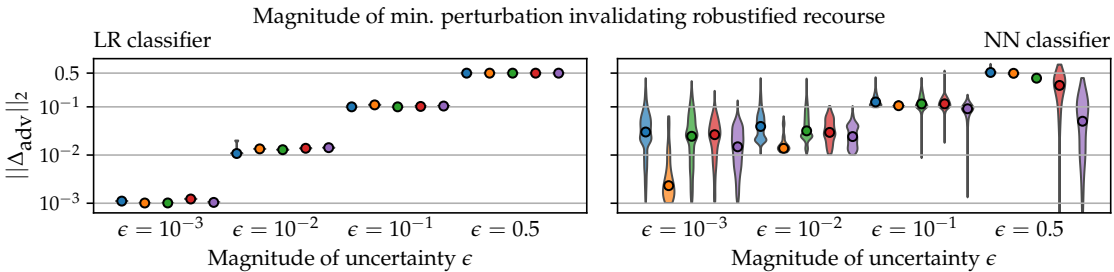


Figure: Legend: COMPAS Adult Loan German Bail.

Conclusion

- Minimum-cost recourse is provably fragile to arbitrarily small changes to x.
- We formalize the notion of adversarial robustness of recourse.
- We present methods to generate adversarially robust recourse.
- We present a model regularizer that facilitates the existence of robust recourse.