On the Adversarial Robustness of Causal Algorithmic Recourse

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

Abstract

Algorithmic recourse seeks to provide actionable recommendations for individuals to overcome unfavorable classification outcomes from automated decision-making systems. Recourse recommendations should ideally be robust to reasonably small uncertainty in the features of the individual seeking recourse. In this work, we formulate the adversarially robust recourse problem and show that recourse methods that offer minimally costly recourse fail to be robust. We then present methods for generating adversarially robust recourse for linear and for differentiable classifiers. Finally, we show that regularizing the decision-making classifier to behave locally linearly and to rely more strongly on actionable features facilitates the existence of adversarially robust recourse.

1. Introduction

Machine learning (ML) classifiers are increasingly being used for consequential decision-making in sensitive domains such as criminal justice and finance (e.g., granting pretrial bail or loan approval). The need to preserve human agency despite the rise in automated decisions faced by individuals has motivated the study of algorithmic recourse, which aims to empower individuals by providing them with actionable recommendations to reverse unfavorable algorithmic decisions (Ustun et al., 2019). Prior works have argued that for recourse to warrant trust, the decision-maker must commit to reversing an unfavorable decision upon the decision-subject fully adopting their prescribed recourse recommendations (Wachter et al., 2017; Venkatasubramanian & Alfano, 2020; Karimi et al., 2022). We argue that if algorithmic recourse is indeed to be treated as a contractual agreement, then recourse recommendations must be robust to plausible uncertainties arising in the recourse process.

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Tübingen, Germany [3]ETH Zürich, Switzerland. Correspondence to: Ricardo Dominguez-Olmedo <ricardo.olmedo@tuebingen.mpg.de>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

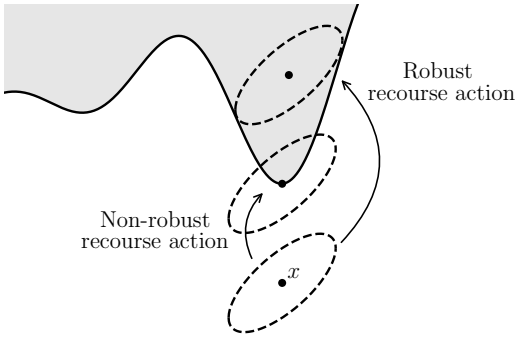

Figure 1: Adversarially robust recourse actions must lead to positive classification outcomes for *all* individuals in the uncertainty set around the individual x seeking recourse.

For instance, consider a banking institution that promises to approve the loan of an individual if they increase their savings by some given amount. Suppose that by the time the individual achieves the prescribed savings increase, the individual's weekly working hours have been slightly reduced due to unforeseen circumstances and the decision-making classifier still deems the individual likely to default on the loan. Shielding recourse against uncertainty *ex-post* by nonetheless granting the loan may be detrimental to both the bank (e.g., monetary loss) and the individual (e.g., bankruptcy and inability to secure future loans), while breaking the recourse promise would negate the effort exerted by the individual and erode trust in the decision maker. We therefore argue for the necessity of ensuring that recourse recommendations are *ex-ante* robust to uncertainty.

In this work, we direct our focus towards robustifying recourse recommendations against uncertainty in the features of the individual seeking recourse. Such uncertainty may arise due to the temporal nature of recourse (e.g., some features may not be static), and/or the presence of noise, adversarial manipulation and other misrepresentations or errors. We adopt a robust optimization view and propose to characterize the uncertainty around the *reported* features of the individual x by defining an uncertainty set $B(x)$ which we assume contains the *true* features of the individual at the time recourse is offered and/or *plausible* future feature values arising from the temporal nature of recourse. We then seek recourse recommendations that remain valid (i.e., lead

to favorable classification outcomes) for *all* plausible individuals in the uncertainty set $B(x)$, as illustrated in Figure 1. We refer to this notion of robustness as the *adversarial robustness of recourse*. We study the adversarial robustness of recourse from the lens of causality (Pearl, 2009). Causal recourse models recourse recommendations as causal interventions on the features of the individual seeking recourse (Karimi et al., 2021), and therefore presents a faithful account of how the features of the individual change as the individual acts on their recourse recommendations, provided that the underlying structural causal model is known or can be approximated from observational data reasonably well (Karimi et al., 2020).

Contributions

- We formulate the adversarially robust recourse problem and show that minimum-cost recourse recommendations are provably fragile to arbitrarily small uncertainty in the features of the individual seeking recourse.
- We present methods for generating adversarially robust causal recourse for linear and for differentiable classifiers. We demonstrate their effectiveness on five tabular datasets, for linear and neural network classifiers.
- We propose a model regularizer that encourages the decision-making classifier to behave locally linearly and to rely more strongly on actionable features. We show that our proposed model regularizer facilitates the existence of adversarially robust recourse.

2. Background and related work

2.1. Background on causality

We assume that the data-generating process of the features $\mathbf{X} = \{X_1, \dots, X_n\}$ of individuals $x \in \mathcal{X}$ is characterized by a known *structural causal model* (SCM) (Pearl, 2009) $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$. The structural equations $\mathbf{S} = \left\{X_i := f_i\left(\mathbf{X}_{\text{pa}(i)}, \mathbf{U}_i\right)\right\}_{i=1}^n$ describe the causal relationship between any given feature X_i, its direct causes $\mathbf{X}_{\text{pa}(i)}$ and some exogenous variable $\mathbf{U}_i$ as a deterministic function f_i. The *exogenous variables* $\mathbf{U} \in \mathcal{U}$, which are distributed according to some probability distribution $P_{\mathbf{U}}$, represent unobserved background factors which are responsible for the variations observed in the data. We assume that the causal graph $\mathcal{G}$ implied by the SCM, with nodes $\mathbf{X} \cup \mathbf{U}$ and edges $\{(v, \mathbf{X}_i) : v \in \mathbf{X}_{\text{pa}(i)} \cup \mathbf{U}_i, i \in [1, n]\}$, is acyclic. The SCM $\mathcal{M}$ then implies a unique *observational distribution* $p_{\mathbf{X}}$ over the features $\mathbf{X}$. Moreover, the structural equations $\mathbf{S}$ induce a mapping $\mathbb{S} : \mathcal{U} \to \mathcal{X}$ between exogenous and endogenous variables. Under the assumption that the exogenous variables are mutually independent (*causal sufficiency*), if there exists some inverse mapping $\mathbb{S}^{-1} : \mathcal{X} \to \mathcal{U}$ such that $\mathbb{S}\left(\mathbb{S}^{-1}(x)\right) = x \ \ \forall x \in \mathcal{X}$, then the endogenous variables corresponding to some individual $x \in \mathcal{X}$ are uniquely identifiable by $\mathbf{U}|x = \mathbb{S}^{-1}(x)$.

SCMs allow for modelling and evaluating the effect of interventions on the system which the SCM models. *Hard interventions* $do(\mathbf{X}_{\mathcal{I}} = \theta)$ (Pearl, 2009) fix the values of a subset $\mathcal{I} \subseteq [d]$ of features $\mathbf{X}_{\mathcal{I}}$ to some $\theta \in \mathbb{R}^{|\mathcal{I}|}$ by altering the structural equations of the intervened upon variables $\mathbf{S}_{\mathcal{I}_i}^{do(\mathbf{X}_{\mathcal{I}}=\theta)} = \mathbf{X}_{\mathcal{I}_i} := \theta_i$ while preserving the rest of the structural equations $\mathbf{S}_i^{do(\mathbf{X}_{\mathcal{I}}=\theta)} = \mathbf{S}_i \ \ \forall i \notin \mathcal{I}$. Thus, hard interventions sever the causal relationship between an intervened upon variable and all of its ancestors in the causal graph. Soft interventions, on the other hand, may modify the structural equations in a more general manner (Korb et al., 2004). In particular, *additive interventions* (Eberhardt & Scheines, 2007) perturb the features $\mathbf{X}$ by some perturbation vector $\Delta \in \mathbb{R}^n$ while preserving all causal relationships, altering the structural equations to $\mathbf{S}^{\Delta} = \left\{X_i := f_i\left(\mathbf{X}_{\text{pa}(i)}, \mathbf{U}_i\right) + \Delta_i\right\}_{i=1}^n$.

Moreover, SCMs imply distributions over *counterfactuals*, allowing to reason about what would have happened under certain hypothetical interventions all else being equal. The counterfactual x^{CF} pertaining to some observed factual individual $x \in \mathcal{X}$ under some hypothetical hard intervention $do(\mathbf{X}_{\mathcal{I}} = \theta)$ (resp. soft intervention Δ) can be computed by first determining the exogenous variables $\mathbf{U}|x = \mathbb{S}^{-1}(x)$ corresponding to the individual x, and then applying the interventional mapping $\mathbb{S}^{do(\mathbf{X}_{\mathcal{I}}=\theta)}$ (resp. $\mathbb{S}^{\Delta}$) from endogenous to exogenous variables (Pearl, 2009). For notational convenience, we denote such mapping as $x^{\text{CF}} = \mathbb{CF}\left(x, do(\mathbf{X}_{\mathcal{I}} = \theta)\right) := \mathbb{S}^{do(\mathbf{X}_{\mathcal{I}}=\theta)}\left(\mathbb{S}^{-1}(x)\right)$ (resp. $x^{\text{CF}} = \mathbb{CF}\left(x, \Delta\right) := \mathbb{S}^{\Delta}\left(\mathbb{S}^{-1}(x)\right)$). We use the notation $x^{\text{CF}} = \mathbb{CF}\left(x, do(\mathbf{X}_{\mathcal{I}} = \theta); \mathcal{M}\right)$ to highlight that the counterfactual x^{CF} follows from a particular SCM $\mathcal{M}$.

2.2. The causal recourse problem

Consider the setting where a classifier $h : \mathcal{X} \to \{0, 1\}$ is used to assign either favorable or unfavorable outcomes to individuals $x \in \mathcal{X}$ (e.g., loan approval). We adopt the causal view of recourse introduced by Karimi et al. (2021) and model recourse recommendations as hard interventions on the features of the individual seeking recourse. We consider interventions of the form $a(x) = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$, where $\theta \in \mathbb{R}^{|\mathcal{I}|}$ is the prescribed change to a subset of the features of the individual x. We consider this additive form, rather than $a = do(\mathbf{X}_{\mathcal{I}} = \theta)$ as Karimi et al. (2021), to explicitly allow for uncertainty in the individual x to propagate to the action $a(x)$. Otherwise, hard-intervening on all features would trivially shield the counterfactual $\mathbb{CF}(x, a)$ from uncertainty in x. For notational simplicity, we refer to an action $a(x)$ as simply a when the pertinent x is clear from context (e.g., we refer to $\mathbb{CF}(x, a(x))$ as simply $\mathbb{CF}(x, a)$).

For a recourse action a to be considered *valid*, the corresponding counterfactual individual must be favorably classified, that is, $h\left(\mathbb{CF}\left(x, a; \mathcal{M}\right)\right) = 1$. Since certain features may be immutable (e.g., race) or bounded (e.g., age), only feasible actions should be recommended. The action feasibility set $\mathcal{F}(x)$ captures the set of feasible actions available to an individual x. Additionally, the cost function $c(x, a)$ models the effort required by an individual $x \in \mathcal{X}$ to implement some recourse action a. Finding the least effortful (i.e., minimum-cost) recourse action for some individual $x \in \mathcal{X}$ amounts to solving the following optimization problem:

$$\begin{aligned} \underset{a(x)=do(\mathbf{X}_{\mathcal{I}}=x_{\mathcal{I}}+\theta)}{\arg\min} \quad & c(x, a) \\ \text{s.t.} \quad & a \in \mathcal{F}(x) \\ & \boxed{h}\left(\mathbb{CF}\left(\boxed{x}, a; \boxed{\mathcal{M}}\right)\right) = 1 \end{aligned} \tag{1}$$

As highlighted in Equation 1, uncertainty in the features of the individual x, the classifier h, and/or the SCM $\mathcal{M}$ may affect the validity of recourse. In Appendix A, we discuss and relate the different sources of uncertainty arising throughout the recourse process.

The non-causal recourse setting is equivalent to the causal recourse setting under the *independently manipulable features* (IMF) assumption, that is, if no causal relationships exist between the features of the individual. Under such assumption, $\mathbb{CF}\left(x, do(\mathbf{X} = x + \theta)\right) = x + \theta$.

2.3. Related work

We now draw connections with existing literature on the robustness of recourse. Previous works have studied the problem of generating recourse actions which remain valid under uncertainty in the classifier h. Pawelczyk et al. (2020) show that recourse actions which place the counterfactual in regions of the feature space with large data support are more robust to competing classifiers that perform equally well. However, recourse actions with large data support may be unnecessarily costly. In contrast, we seek robust recourse actions that are also minimally costly. Another line of work has considered robustness of recourse against changes to the classifier in response to dataset shift. Rawal et al. (2020b) show that recourse actions are typically not robust to such model changes, and Upadhyay et al. (2021) aim to mitigate this issue by generating recourse with a minimax optimization procedure where the cost of recourse is minimized subject to the recourse action being valid under adversarial changes to the classifier h. While we adopt a similar minimax approach to generate robust recourse, we focus on robustifying recourse against uncertainty in the individual x rather than the classifier h. Subsequent works propose to instead adopt a distributionally robust viewpoint (Black et al., 2022; Bui et al., 2022; Guo et al., 2022). Likewise, a natural extension of our work is to adopt a distributionally robust optimization viewpoint.

Regarding robustness of recourse against uncertainty in the SCM $\mathcal{M}$, Karimi et al. (2020) consider the setting where the underlying SCM is not know and thus must be approximated from data, and propose a method to generate recourse recommendations which have low probability of being invalid due to the misspecification of the underlying SCM. Our work is tangential to Karimi et al. (2020).

Previous works have identified that small changes to the features of the decision-subject x may result in different recourse recommendations being offered with potentially very different costs (von Kügelgen et al., 2022; Slack et al., 2021; Artelt et al., 2021). Instead of focusing on the cost of recourse, we study whether recourse actions remain valid under uncertainty in the individual x. In Appendix D, we discuss in more detail the relation between these two different notions of robustness. The concurrent work of Virgolin & Fracaros (2022) is most similar to ours, as they study the robustness of recourse to adversarial perturbations to the individual x. They propose an evolutionary algorithm to generate robust recourse, which they evaluate for random forest classifiers. In contrast, we focus on generating recourse for differentiable classifiers, in particular linear and neural network models. Additionally, we consider the more general causal recourse setting, and we model feature perturbations in a causal manner. Lastly, Pawelczyk et al. (2022) study the rate at which random perturbations to the actionable features of x invalidate recourse, and propose a method to generate recourse which is less likely to be invalidated. In contrast, we aim to generate recourse which is, to the extent possible, provably robust against adversarial perturbations. We additionally consider the causal setting.

Finally, Ross et al. (2021) propose to regularize the decision-making classifier at training time to facilitate the existence of recourse. They propose a regularizer which encourages the classifier to be very sensitive (i.e., fragile) to changes to the actionable features. Such fragility of the classifier, however, can be problematic when generating adversarially robust recourse. In contrast, our proposed regularizer encourages the classifier to be more robust to adversarial examples by regularizing it to behave locally linearly (Qin et al., 2019).

3. Counterfactual uncertainty sets

In the adversarial robustness literature, the uncertainty surrounding an observation x is often modelled by an ϵ-ball of uncertainty $B(x) = \{x + \Delta | \, \|\Delta\| \leq \epsilon\}$ around the observed x, where the norm $\|\cdot\|$ characterizes some relevant notion of magnitude for perturbations Δ to the observation x and ϵ specifies the amount of uncertainty under consideration (Madry et al., 2018; Bertsimas et al., 2019). Intuitively,

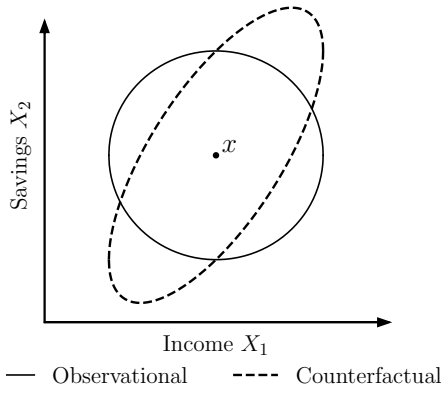

Figure 2: Illustration of the observational and counterfactual neighborhoods of similar individuals for the linear SCM $X_1 = U_1$ (Income), $X_2 = X_1 + U_2$ (Savings) under $\|\cdot\|_2$.

small perturbations Δ to the observation x result in plausibly similar data points. Then, the uncertainty set $B(x)$ amounts to an ϵ-neighborhood of plausible data points similar to x.

From a causal perspective, an underlying assumption in the above definition of $B(x)$ is that the features are independently manipulable. We argue, however, that explicitly considering the causal relationships between features can result in potentially more informative neighborhoods of similar individuals. To account for the causal effects of the perturbations Δ, we propose to model such perturbations as additive interventions on the features of the individual x.

Definition 1 (Counterfactual neighborhood of similar individuals)**.** *For some individual x, SCM $\mathcal{M}$ and norm $\|\cdot\|$, we define the counterfactual ϵ-neighborhood $B(x)$ of individuals similar to the observed x as the set of causal counterfactuals*[1] *under all possible ϵ-small additive interventions:*

$$B(x) = \{\mathbb{CF}(x, \Delta; \mathcal{M}) \mid \|\Delta\| \leq \epsilon\} \tag{2}$$

As a motivating example, consider the SCM $\mathcal{M}$ with features $X_1 = U_1$ and $X_2 = X_1 + U_2$ respectively denoting the income and savings of some individual x. Figure 2 illustrates the corresponding observational and counterfactual neighborhoods $B(x)$. Observe that the counterfactual neighborhood is skewed towards individuals with both higher (or lower) income and savings, since we would expect perturbations that simultaneously decrease income and increasing savings to be less likely to be observed (i.e., are not well-supported by the causal structure of the data). Therefore, we argue that counterfactual neighborhoods can be more informative, since the causal relationships between features are explicitly considered. Additionally, for non-linear SCMs, counterfactual neighborhoods $B(x)$ adapt to the local geometric structure of the data manifold (see Appendix B).

[1] As introduced in Section 2.1, do not confuse with the notion of "counterfactual examples" from the recourse literature.

4. The adversarially robust recourse problem

We consider the problem of generating recourse actions that are robust to uncertainty in the features of the individual seeking recourse. We adopt a robust optimization viewpoint (Ben-Tal et al., 2009) and model uncertainty in the individual x by characterizing an uncertainty set $B(x)$ of plausible individuals around x. We formalize the notion of adversarial robustness of recourse as follows:

Definition 2 (Adversarially robust recourse action)**.** *For some classifier h, individual $x \in \mathcal{X}$, and uncertainty set $B(x)$, a recourse action a is adversarially robust if it is valid for all individuals in the uncertainty set $B(x)$*

$$h(\mathbb{CF}(x', a)) = 1 \quad \forall x' \in B(x) \tag{3}$$

We additionally define the adversarially robust recourse problem by directly incorporating the above robustness requirement as a constraint to the standard recourse problem:

Definition 3 (Adversarially robust recourse problem)**.** *For some uncertainty set $B(x)$, the minimum-cost recourse action which is adversarially robust is given by*

$$\begin{aligned} \underset{a(x)=do(\mathbf{X}_\mathcal{I}=x_\mathcal{I}+\theta)}{\arg\min} \quad & c(x, a) \\ \textit{s.t.} \quad & a \in \mathcal{F}(x) \\ & h(\mathbb{CF}(x', a)) = 1 \;\; \forall x' \in B(x) \end{aligned} \tag{4}$$

Note that any feasible solution to the above optimization problem is by definition adversarially robust.

4.1. Recourse is fragile under mild conditions

We show that under mild conditions on the cost function c, feasibility set $\mathcal{F}(x)$ and SCM $\mathcal{M}$, minimum-cost recourse actions are provably fragile to arbitrarily small uncertainty in the features of the individual seeking recourse.

Theorem 1. *Let a^* be the solution to the standard recourse optimization problem defined in Equation 1. If*

(i) The cost function $c(x, do(X_\mathcal{I} = x_\mathcal{I} + \theta))$ is strictly convex in θ with minimum $\theta = 0$
(ii) $do(X_\mathcal{I} = x_\mathcal{I} + \theta) \in \mathcal{F}(x) \implies do(X_\mathcal{I} = x_\mathcal{I} + t\theta) \in \mathcal{F}(x) \quad \forall\, 0 < t < 1$
(iii) The SCM $\mathcal{M}$ is an additive noise model (Pearl, 2009).

Then, for any $\epsilon > 0$ there exists some plausible individual $x' \in B(x)$ in the ϵ-neighbourhood of similar individuals $B(x) = \{\mathbb{CF}(x, \Delta) \,|\, \|\Delta\| \leq \epsilon\}$ such that a^* is not a valid recourse action, i.e., $h(\mathbb{CF}(x', a^*)) = 0$. That is, the minimum-cost recourse action a^* is fragile to arbitrarily small uncertainty in the individual x seeking recourse.

Condition (i) requires that larger feature changes imply strictly more effort. This condition is satisfied by weighted

p-norms (Karimi et al., 2020) and percentile costs (Ustun et al., 2019), the most widely used cost functions in the recourse literature. Condition (ii) states that if it is feasible to change a feature by some amount, then it must also be feasible to change that feature to a lesser degree. This condition is satisfied by box actionability constrains, commonly assumed in the recourse literature (Karimi et al., 2022). Lastly, condition (iii) is a common modelling assumption for estimating the SCM $\mathcal{M}$ from data (Karimi et al., 2020), and also holds in the non-causal recourse setting.

Therefore, in the settings commonly considered by the algorithmic recourse literature, methods seeking minimum-cost recourse offer provably fragile recourse recommendations.

4.2. On the existence of adversarially robust recourse

We extend the sufficient conditions for the existence of recourse derived by Ustun et al. (2019) to the adversarially robust causal recourse setting. First, we state the negative result that even if recourse exists for every individual $x \in \mathcal{X}$, robust recourse may not exist for any individual $x \in \mathcal{X}$.

Proposition 1. *The existence of recourse for all individuals $x \in \mathcal{X}$ is not a sufficient condition for the existence of adversarially robust recourse for any $x \in \mathcal{X}$, even under the strong assumption that all features are actionable.*

Intuitively, for robust recourse to exist the decision-making classifier must be minimally robust in the sense that there must exist at least one individual which is robustly classified. In this sense, robustness of prediction is necessary (but not sufficient) for the existence of robust recourse. Under mild robustness conditions on the classifier, all features being actionable is sufficient for the existence of robust recourse.

Proposition 2. *For an SCM $\mathcal{M}$ with linear structural equations, if all features are actionable and there exists some robustly classified individual $x^+ \in \mathcal{X}$ such that $h(x') = 1 \ \forall x' \in B(x^+)$, then there exists an adversarially robust recourse action for every individual $x \in \mathcal{X}$.*

Intuitively, one can then require all negatively classified individuals to act such that their features resemble x^+. Note that the above statement holds only for linear SCMs, since for non-linear SCMs the shape of the uncertainty set $B(x)$ differs between individuals x, and thus it is not possible to state a single robustness requirement for h at x^+.

By further assuming that the classifier h is linear, it is possible to relax the condition of all features being actionable to a single feature being actionable and unbounded:

Proposition 3. *For a linear classifier h and linear SCM $\mathcal{M}$, under mild conditions on the weights of the classifier described in Appendix C.4, if there exists a feature $\mathbf{X}_j$ that is actionable and unbounded, then there exists an adversarially robust recourse action for every $x \in \mathcal{X}$.*

5. Generating adversarially robust recourse

5.1. The linear case

For a linear classifier $h(x) = \langle w, x \rangle \geq b$ and linear SCM, we show that generating robust recourse for the classifier h is equivalent to generating standard recourse for a modified linear classifier $h'(x) = \langle w, x \rangle \geq b'$ whose "acceptance threshold" is sufficiently increased (i.e., $b' \geq b$).

Proposition 4. *Let $h(x) = \langle w, x \rangle \geq b$ be a linear classifier, $\mathcal{M}$ an SCM with linear structural equations, and $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$ an uncertainty set of plausible individuals. Then, an action $a(x) = do(\mathbf{X}_\mathcal{I} = x_\mathcal{I} + \theta)$ is an adversarially robust recourse action if and only a is a valid recourse action for the following modified classifier:*

$$h'(x) = \langle w, x \rangle \geq b + \left\| J_{\mathbb{S}^\mathcal{I}}^T w \right\|^* \epsilon \tag{5}$$

where $\|\cdot\|^*$ denotes the dual norm of $\|\cdot\|$ and $J_{\mathbb{S}^\mathcal{I}}$ denotes the Jacobian of the interventional mapping resulting from hard-intervening on a subset of features $\mathbf{X}_\mathcal{I}$.

Corollary 1. *For any given individual $x \in \mathcal{X}$, the minimum-cost adversarially robust recourse action for the original classifier h is equivalent to the minimum-cost standard recourse action for the modified classifier h'.*

Therefore, in the linear setting, per Corollary 1 any method for generating standard recourse can be readily used to generate adversarially robust recourse by simply considering the modified classifier h'. In such cases, adversarial robustness can be straightforwardly embedded within methods seeking to promote other desiderata, such as large data-support (Joshi et al., 2019) or fairness constraints (Gupta et al., 2019; von Kügelgen et al., 2022).

5.2. The differentiable case

We now consider the setting where the classifier h and SCM $\mathcal{M}$ are differentiable. First, we derive an unconstrained penalty problem that is equivalent to the adversarially robust recourse problem. We then discuss how to efficiently solve said unconstrained optimization problem.

Proposition 5. *Let $h(x) = \tilde{h}(x) \geq b$ for differentiable $\tilde{h} : \mathcal{X} \to [0, 1]$. The adversarially robust recourse problem is equivalent to the following unconstrained problem:*

$$\min_{a \in \mathcal{F}(x)} \max_{\lambda \geq 0} c(x, a) + \lambda \left(\log b + \max_{x' \in B(x)} \ell \left(\tilde{h} \left(\mathbb{CF}(x', a) \right), \mathbf{1} \right) \right)$$

where ℓ is the binary cross-entropy loss.

To solve the outer minimax optimization problem, we adopt the causal recourse approach of Karimi et al. (2020) and use projected gradient descent over the recourse action a and feasibility set $\mathcal{F}(x)$, while iteratively increasing λ to place growing emphasis in crossing the classifier's decision

Algorithm 1 Generate adversarially robust recourse for a differentiable classifier h and differentiable SCM $\mathcal{M}$.

Require: Factual individual x, uncertainty set $B(x)$, subset $\mathcal{I}$ of intervened-upon features $x_{\mathcal{I}}$, $\lambda > 0, \gamma > 1, \theta = 0$
while $N \leq N_{\max}$ **do**
  while not converged **do**
    $a(x) \leftarrow do(X_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$
    $x^* \leftarrow \arg\max_{x' \in B(x)} \ell\left(\tilde{h}(\mathbb{CF}(x', a)), \mathbf{1}\right)$
    if $h\left(\mathbb{CF}\left(x^*, a\right)\right) = 1$ **then**
      return θ
    $g \leftarrow \nabla_\theta \left(c(x, a) + \lambda\, \ell\left(\tilde{h}(\mathbb{CF}(x^*, a)), \mathbf{1}\right)\right)$
    $\theta \leftarrow \text{Proj}_{\mathcal{F}(x)}(\theta - \alpha g)$
  $\lambda \leftarrow \gamma\lambda$

boundary, as described in Algorithm 1. Note that if there is no uncertainty in the features of the individual x, that is $B(x) = \{x\}$, then the optimization procedure is precisely that of Karimi et al. (2020) and Wachter et al. (2017) for the causal and non-casual recourse settings respectively, since the solution to the inner maximization is trivially $x^* = x$.

We solve the inner maximization in Proposition 5 using projected gradient ascent over the uncertainty set $B(x)$. Since this inner maximization problem is in general non-convex, the local maximum x^* found with gradient ascent may not be the global maximum in $B(x)$. If x^* is not the global maximum, then in Algorithm 1 the exit condition $h\left(\mathbb{CF}\left(x^*, a\right)\right) = 1$ does not imply that $h\left(\mathbb{CF}\left(x', a\right)\right) = 1\ \forall x' \in B(x)$. Thus, in general it is not possible to guarantee that the recourse actions returned by the proposed algorithm are adversarially robust. However, as discussed in the experiments section, we empirically find that, for sufficiently small uncertainty ϵ, the proposed algorithm is effective in generating robust recourse actions.

6. Actionability regularization

While the robustness of recourse is desirable for both the decision-maker and the decision-subject, the burden of immunizing recourse against uncertainty falls solely on the decision-subject: robust recourse is more effortful (minimum-cost recourse is provably fragile) and may not even exist (Proposition 1). To regulate the burden of robustness between the decision-maker and the decision-subject, we propose to regularize the decision-making classifier at training time with the aim of 1) facilitating the existence of robust recourse, and 2) reducing the extra cost of seeking robust recourse. Regularizing the classifier to promote such desiderata may come at a cost in predictive performance, thus shifting part of the burden of robust recourse from the decision-subject to the decision-maker. We propose to regularize the decision-making classifier to behave locally linearly and to rely more strongly on actionable features.

6.1. Theoretical motivation

To motivate our proposed regularizer, we refer to the sufficient conditions for the existence of adversarially robust recourse presented in Section 4.2. Per Proposition 2 and Proposition 3, we hypothesize, respectively, that using only actionable features for classification and using a linear classifier facilitates the existence of robust recourse. However, often times most available features are unactionable, and thus classifiers trained only with actionable features will exhibit poor predictive performance. Similarly, the predictive performance of linear classifiers is generally inferior to that of nonlinear classifiers. As a middle ground, we propose to use all available features to train a potentially non-linear classifier, but regularize the classifier to rely more strongly on the actionable features and to locally behave linearly.

Such choice of regularization is additionally well-motivated from the viewpoint of reducing the extra cost of robustifying recourse, for which we derive the following upper bound:

Proposition 6. *Let $h(x)$ be a linear classifier $\langle w, x\rangle \geq b$, $x \in \mathcal{X}$ a negatively classified individual for which there exists some recourse action $a(x) = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$, and $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$ the uncertainty set. If the features $X_{\mathcal{I}}$ are unbounded and independently manipulable, and the cost function is subadditive, then there exists some adversarially robust recourse action a' such that the extra cost of robustifying recourse is upper bounded by*

$$\frac{c(x, a') - c(x, a)}{c(x, a)} \leq \frac{\|m_{\mathcal{A}} \odot w\|^* + \|m_{\tilde{\mathcal{A}}} \odot w\|^*}{\langle m_{\mathcal{A}} \odot w, \theta\rangle}\epsilon \quad (6)$$

where $m_{\mathcal{A}} \in [0, 1]^n$ (resp. $m_{\tilde{\mathcal{A}}}$) is the mask vector for the set of actionable features $\mathcal{A}$ (resp. unactionable features $\tilde{\mathcal{A}}$), and $\|\cdot\|^*$ is the dual norm of $\|\cdot\|$.

Observe that the upper bound presented above is reduced if the unactionable features are less discriminative (i.e., if $\|m_{\tilde{\mathcal{A}}} \odot w\|^*$ is small). By additionally regularizing non-linear classifiers to behave locally linearly, we ensure that this upper bounds approximately holds locally. In Appendix C.7, we extend the above upper bound to the causal setting.

6.2. The Actionable Locally Linear Regularizer

To formalize our proposed regularizer, we draw inspiration from local linearity regularization (Qin et al., 2019), a popular regularization technique from the adversarial robustness literature. For differentiable classifiers $h(x)$, we propose the Actionable Locally Linear Regularizer (ALLR):

$$\begin{aligned}\mathcal{R}(x) = \mu_1 \max_{\|\delta\| \leq \epsilon} |h(x + \delta) - \langle \delta, \nabla_x h(x)\rangle - h(x)| \\ + \mu_2 \|m_{\tilde{\mathcal{A}}} \odot \nabla_x h(x)\|\end{aligned} \quad (7)$$

The first term in the ALLR regularizer encourages the classifier h to behave linearly near the observed data, while the

second term encourages unactionable features to not be very discriminative. The hyperparameters $\mu_1, \mu_2 \in \mathbb{R}$ determine the strength of regularization. The classifier is then trained using regularized risk minimization:

$$\min_{\psi} \mathbb{E}_{(x,y)\sim p(x,y)} \left[\ell(h_\psi(x), y) + \mathcal{R}(x)\right] \tag{8}$$

where ℓ is the binary cross-entropy loss, $p(x, y)$ is the training data distribution, and ψ are the weights of the classifier.

7. Experiments and results

First, we empirically validate the effectiveness of the proposed methods in generating adversarially robust recourse. Secondly, we empirically show that regularizing the decision-making classifier with the proposed ALLR regularizer facilitates finding adversarially robust recourse actions. We open source our implementations and experiments[2].

We consider four real-world datasets and one semi-synthetic dataset. For the causal recourse setting, we consider the COMPAS recidivism dataset (Larson et al., 2016) and the Adult demographic dataset (Kohavi & Becker, 1996), for which we adopt the causal graphs assumed in Nabi & Shpitser (2018). We fit the structural equations using linear models and neural nets for the linear and the non-linear case, respectively. We additionally consider one semi-synthetic SCM introduced by Karimi et al. (2020), which is inspired in a loan approval setting. For the non-causal recourse setting, we consider the South German Credit dataset (Groemping, 2019), as well as a recidivism dataset (Schmidt & Witte, 1988) from North Carolina which we refer to as Bail. In Appendix E.1.1 we list the actionability constraints considered.

For all datasets, we treat actionable categorical variables as real-valued, and we standardize all real-valued features. We use as the cost function the ℓ_1 norm, that is $c(x, a) = \|\theta\|_1$ for $a(x) = do(X_\mathcal{I} = x_\mathcal{I} + \theta)$. We consider two types of classifiers: linear models trained with logistic regression (LR), and neural network (NN) models. We define the uncertainty set $B(x)$ with respect to the 2-norm. Since features are standardized, robustifying against ϵ uncertainty is equivalent to guarding against perturbations at least ϵ times the standard deviation of any given feature. For instance, in the Adult dataset the standard deviation of the feature *age* is 13.6 years. Thus, for $\epsilon = 0.1$ uncertainty robust actions should remain valid even if the age of the individual seeking recourse changes by $\pm$ 1.36 years. In Appendix E.1.2 we list the standard deviation of the considered features.

7.1. Minimum-cost recourse is fragile

First, we empirically demonstrate that recourse methods which aim to generate minimum-cost recourse fail to be robust. To do so, we train the decision-making classifiers using expected risk minimization and generate recourse for the negatively classified individuals. We use the minimum-cost recourse methods of Wachter et al. (2017) and Karimi et al. (2021) for the causal and non-causal recourse setting, respectively. We then use the C&W adversarial attack (Carlini & Wagner, 2017) to find the smallest additive intervention Δ_{adv} which invalidates the generated recourse action a, that is, such that $h(\mathbb{CF}(x_{\text{adv}}, a)) = 0$ for $x_{\text{adv}} = \mathbb{CF}(x, \Delta_{\text{adv}})$.

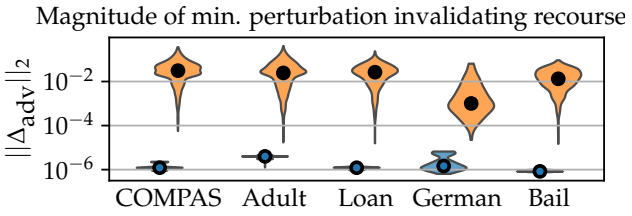

Figure 3: Fragility of standard recourse. Very small feature perturbations can invalidate recourse, particularly for linear classifiers. Legend: ■ LR classifier ■ NN classifier.

We present the experiment results in Figure 3. We observe that the recourse actions generated for both LR and NN classifiers are remarkably fragile, and can be invalidated under uncertainty ranging from $\epsilon = 10^{-2}$ to 10^{-6}. For linear classifiers, recourse actions are particularly brittle, since the recourse problem is convex and thus its global minimum (which is provably fragile) can be accurately computed.

7.2. Generating adversarially robust recourse

We evaluate the effectiveness of the methods proposed in Section 5 in generating adversarially robust recourse against different uncertainty $\epsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 0.5\}$. We consider $\epsilon \in \{10^{-3}, 10^{-2}\}$ to be reasonably small uncertainty, $\epsilon = 0.1$ significant uncertainty, and $\epsilon = 0.5$ very large uncertainty. For each individual x and level of uncertainty ϵ, we first generate a recourse action a which is robustified against ϵ. We then use the C&W attack to search for the smallest intervention Δ_{adv} to the features of the individual x which invalidates the generated recourse action a. If the magnitude of such adversarial intervention is lower than the uncertainty ϵ against which the a was robustified (i.e., $\|\Delta_{\text{adv}}\| \leq \epsilon$), we can assert that the action a is fragile.

We present the experimental results in Figure 4. For linear models, all adversarial interventions Δ_{adv} found have magnitude larger than ϵ. Thus, in the linear case our proposed method is effective in generating robust recourse. Furthermore, since all perturbations found are larger than ϵ by only an arbitrarily small amount, the recourse actions generated are not only robust but also minimally costly (i.e., there is no over-robustification). For NNs models, our proposed method is effective in generating robust recourse actions against reasonably small uncertainty ϵ. For reasonably large

[2] github.com/RicardoDominguez/AdversariallyRobustRecourse

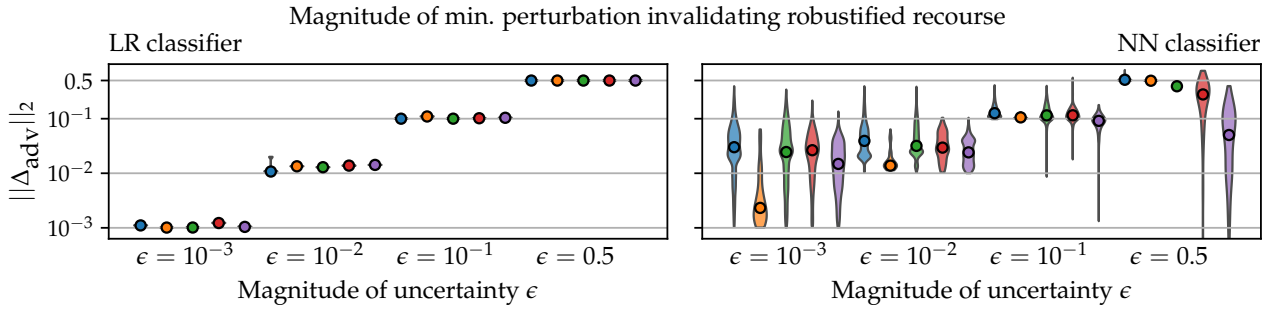

Figure 4: Robustness of recourse actions that have been robustified against ϵ uncertainty. For LR classifiers, the proposed methods are very effective in generating robust recourse actions. For NN classifiers, some of the generated recourse actions may be fragile for reasonably large uncertainty $\epsilon \in \{0.1, 0.5\}$. Legend: COMPAS Adult Loan German Bail.

uncertainty ϵ, however, some of the generated recourse actions may be fragile, since the non-covex inner maximization in Algorithm 1 is more likely to arrive to a local maximum rather than to its global maximum. Nonetheless, our proposed method generates substantially less brittle recourse actions compared to the standard minimum-cost recourse generation methods previously evaluated (Section 7.1).

7.3. Actionable local linearity regularization

We empirically evaluate whether training the decision-making classifier with the proposed ALLR regularizer facilitates the existence of adversarially robust recourse. We compare against the following approaches:

- Empirical risk minimization (ERM): the standard choice for model training in the recourse literature. Equivalent to ALLR with zero regularization strength.
- ERM using only actionable features (AF): facilitates the existence of recourse per Proposition 2. Equivalent to ALLR with infinitely large regularization $\mu_2 \to \infty$.
- The approach of Ross et al. (2021), which makes actionable features more discriminative using the regularizer $\mathcal{R}(x) = \min_\delta \ell(h(x + m_\mathcal{A} \odot d), \mathbf{1})$.

First, we train five classifiers with different train and test splits for each of the considered approaches. For the negatively classified individuals, we generate standard recourse (i.e., $\epsilon = 0$) and robust recourse against a significant amount of uncertainty $\epsilon = 0.1$. We then evaluate the percentage of individuals for which robust recourse is found, as well as the mean cost of recourse. We additionally evaluate the predictive performance of the classifiers by computing their accuracy and Matthews correlation coefficient (MCC).

We present the experimental results in Figure 5. For both LR and NN models, we find that our proposed regularizer is generally very effective in facilitating the existence of adversarially robust recourse, increasing the percentage of individuals for which robust recourse is found by up to 100%. Furthermore, classifiers trained with ALLR generally offer recourse at a similar or lower cost compared to classifiers trained with ERM. The accuracy of the classifiers trained with ALLR may decrease by up to 3% compared to classifiers trained with ERM, but does not decrease at all for three out of the five datasets considered. Using only actionable features (AF) for classification also significantly facilitates the existence of robust recourse, but often leads to poor predictive performance. Finally, while the regularization approach of Ross et al. (2021) is very effective in providing low-cost recourse, ALLR is generally more effective in facilitating the existence of robust recourse.

8. Conclusion

Uncertainty in the recourse process is inevitable. Previously suggested *ex-post* solutions to mitigate the effect of uncertainty in the recourse process may result in negative outcomes for both the decision-maker and the individual. We instead adopted an *ex-anti* approach to the robustness of recourse by requiring recourse recommendations to be robust to uncertainty in the features of the individual seeking recourse. Alarmingly, we showed that minimum cost recourse is provably fragile to arbitrarily small uncertainty in the individual seeking recourse. To address this critical issue, we formulated the adversarially robust recourse problem and presented methods to generate adversarially robust recourse for linear and for differentiable classifiers. Finally, we derived sufficient conditions for the existence of adversarially robust recourse, and we empirically demonstrated that regularizing the decision-making classifier to behave locally linearly and to rely more strongly on actionable features facilitates the existence of robust recourse.

Acknowledgements The authors thank Shalmali Joshi, Himabindu Lakkaraju, Martin Pawelczyk and Berk Ustun for helpful feedback and discussions. AHK acknowledges generous founding support from NSERC, CLS, and Google.

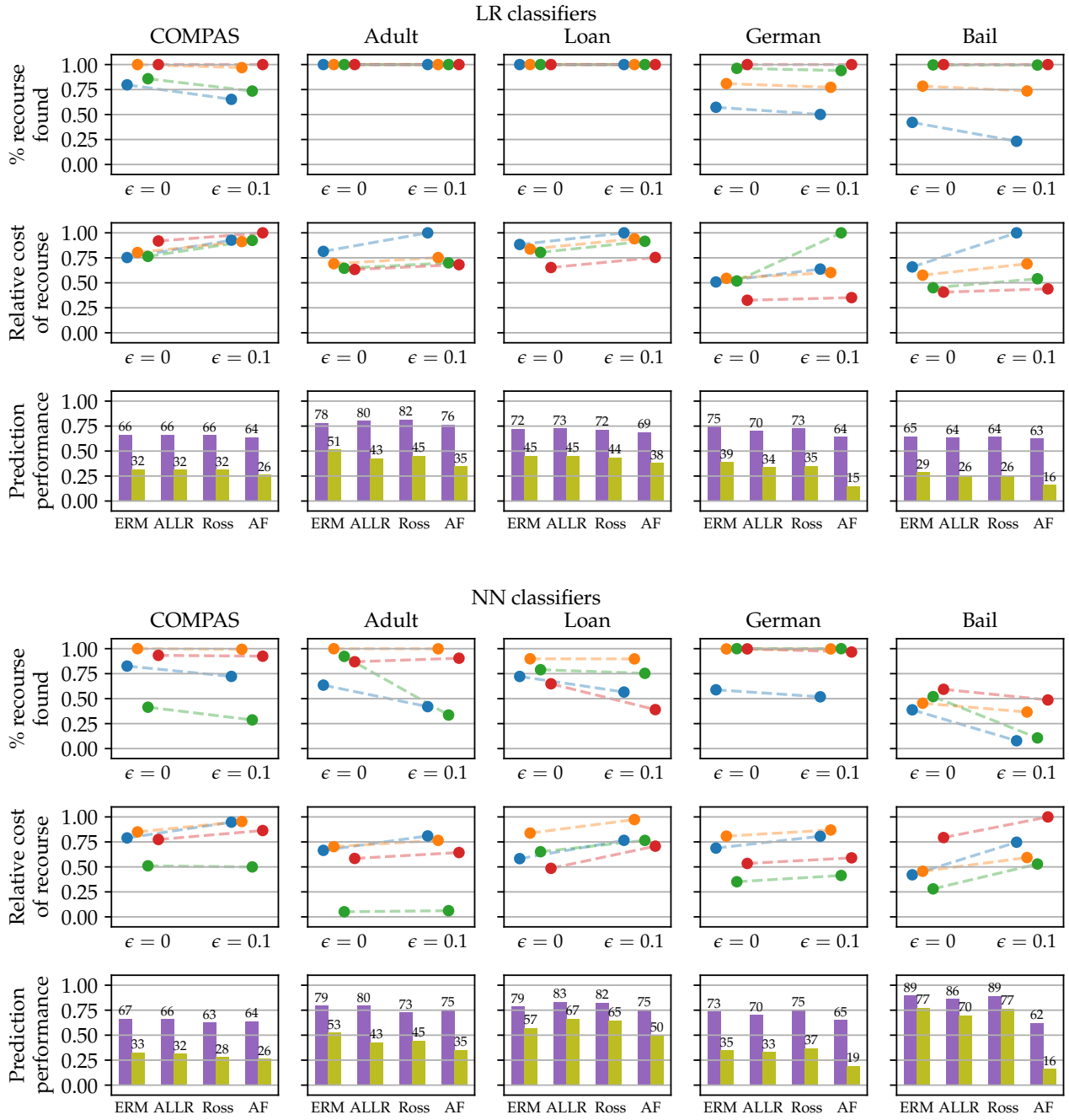

Figure 5: For LR models, classifiers trained with ALLR provide robust recourse to a substantially larger percentage of individuals compared to classifiers trained using ERM. For NN models, ALLR is the most effective method in facilitating the existence of adversarially robust recourse. Legend: ERM ALLR Ross et al. AF Accuracy MCC.

References

Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., and Hammer, B. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 01–09. IEEE, 2021.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*. Princeton university press, 2009.

Bertsimas, D., Dunn, J., Pawlowski, C., and Zhuo, Y. D. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

Black, E., Wang, Z., Fredrikson, M., and Datta, A. Consistent counterfactuals for deep models. *International Conference on Learning Representations*, 2022.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Breiman, L. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Bui, N., Nguyen, D., and Nguyen, V. A. Counterfactual plans under distributional ambiguity. 2022.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. *Advances in Neural Information Processing Systems*, 29: 1632–1640, 2016.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Groemping, U. South german credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4:2019, 2019.

Guo, H., Jia, F., Chen, J., Squicciarini, A., and Yadav, A. Rocoursenet: Distributionally robust training of a prediction aware recourse model. *arXiv preprint arXiv:2206.00700*, 2022.

Gupta, V., Nokhiz, P., Roy, C. D., and Venkatasubramanian, S. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *SafeML Workshop at the International Conference on Learning Representations (ICLR)*, 2019.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, pp. 265–277, 2020.

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *ACM Computing Surveys (CSUR)*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Kohavi, R. and Becker, B. Uci adult data set. *UCI Meachine Learning Repository*, 6, 1996.

Korb, K. B., Hope, L. R., Nicholson, A. E., and Axnick, K. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 322–331. Springer, 2004.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Murphy, P. M. Uci repository of machine learning databases. *ftp:/pub/machine-learning-databaseonics.uci.edu*, 1994.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Pawelczyk, M., Broelemann, K., and Kasneci, G. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

Pawelczyk, M., Datta, T., van-den Heuvel, J., Kasneci, G., and Lakkaraju, H. Algorithmic recourse in the face of noisy human responses. *ICLR 2022 Workshop on Socially Responsible Machine Learning*, 2022.

Pearl, J. *Causality*. Cambridge university press, 2009.

Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32:13847–13856, 2019.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. *Dataset shift in machine learning*. Mit Press, 2009.

Rawal, K., Kamar, E., and Lakkaraju, H. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020a.

Rawal, K., Kamar, E., and Lakkaraju, H. Can i still trust you?: Understanding the impact of distribution shifts on algorithmic recourses. *arXiv preprint arXiv:2012.11788*, 2020b.

Ross, A., Lakkaraju, H., and Bastani, O. Learning models for actionable recourse. *Advances in Neural Information Processing Systems*, 34, 2021.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Schmidt, P. and Witte, A. D. *Predicting recidivism in north carolina, 1978 and 1980*. Inter-university Consortium for Political and Social Research, 1988.

Slack, D., Hilgard, S., Lakkaraju, H., and Singh, S. Counterfactual explanations can be manipulated. *35th Conference on Neural Information Processing Systems*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards robust and reliable algorithmic recourse. *35th Conference on Neural Information Processing Systems*, 2021.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.

Venkatasubramanian, S. and Alfano, M. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 284–293, 2020.

Virgolin, M. and Fracaros, S. On the robustness of counterfactual explanations to adverse perturbations. *arXiv preprint arXiv:2201.09051*, 2022.

von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. On the fairness of causal algorithmic recourse. *AAAI Conference on Artificial Intelligence*, 2022.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

A. Uncertainties in the recourse process

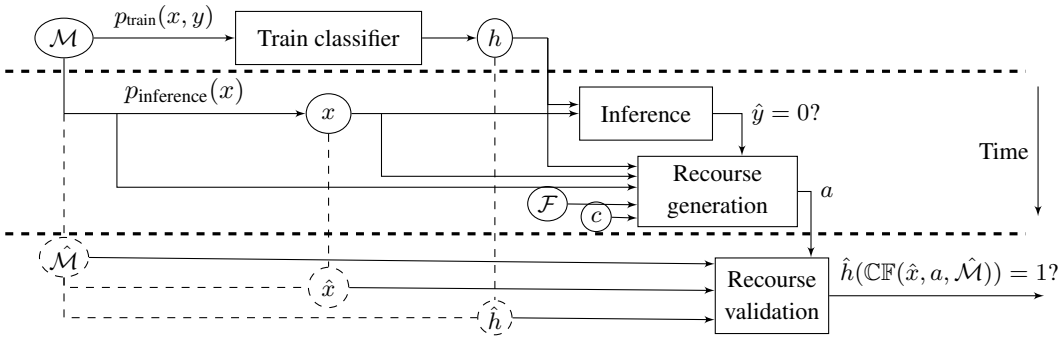

Figure 6: Overview of the recourse process. Uncertain elements are represented with dashed circles. Possible relations between uncertain elements are represented with non-bold dashed lines. Bold dashed lines represent temporal jumps.

Uncertainties may arise throughout the recourse process, as depicted in Figure 6. Some well-studied sources of uncertainty in the classification setting naturally extend to algorithmic recourse. A great deal of the robust classification literature has focused on uncertainty in the inputs x at inference time, which may arise due to the presence of noise (Fawzi et al., 2016; Xu et al., 2009), adversarial manipulation (Madry et al., 2018; Szegedy et al., 2014) and other misrepresentations or errors in the data (Zheng et al., 2016). Regarding the classifier h, the optimization problem solved during model training often does not have a unique optimal solution and multiple models may perform equally well (Breiman, 2001; Rudin, 2019). Moreover, the temporal nature of recourse introduces a unique challenge: the circumstances under which recourse is generated may change by the time the individual is able to implement their prescribed recourse action. For instance, the distribution over inputs itself may change at inference time, under phenomena such as data-set shift (Moreno-Torres et al., 2012; Quiñonero-Candela et al., 2009) or for tasks requiring out of distribution generalization (Geirhos et al., 2020; Muandet et al., 2013).

From a causal perspective, changes in the observational data distribution are a consequence of changes to the underlying SCM (Bühlmann, 2020). Indeed, the data-generation process characterized by the SCM $\mathcal{M}$ may be imperfectly known (von Kügelgen et al., 2022) or may dynamically change over time to some other SCM $\hat{\mathcal{M}}$. Consequently, the counterfactual individual resulting from the prescribed recourse intervention may also change. Furthermore, decision-makers may have to periodically retrain their models to prevent performance degradation due to the distribution shift resulting from a change in the SCM, producing further uncertainty over the future classifier $\hat{h}$ (Rawal et al., 2020a; Upadhyay et al., 2021). Finally, it may be unreasonable to expect the individual x to not suffer changes outside its control over an extended period of time (Venkatasubramanian & Alfano, 2020), leading to uncertainty in the future individual $\hat{x}$. Thus, acting on the prescribed recourse may not lead to favorable classification due to changes to the SCM $\hat{\mathcal{M}}$, classifier $\hat{h}$, and/or factual individual $\hat{x}$.

B. Counterfactual neighborhoods adapt to the local geometry of the data manifold

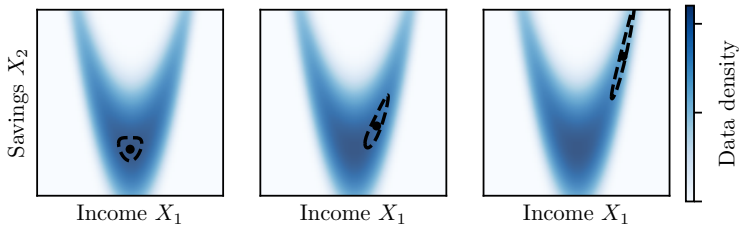

Figure 7: Counterfactual neighborhoods of similar individuals for the SCM $X_1 = U_1$, $X_2 = X_1^2 + U_2$.

The SCM $\mathcal{M}$ amounts to a generative model of the data and thus (approximately) captures the underlying data manifold. Therefore, the individuals in the uncertainty set $B(x)$ are realistic in the sense that they lie within the data manifold (i.e., have sufficient data support). For non-linear SCMs, the shape of $B(x)$ adapts to the local geometry of the manifold, as illustrated in Figure 7. Further studying this behavior is an interesting research direction for future work.

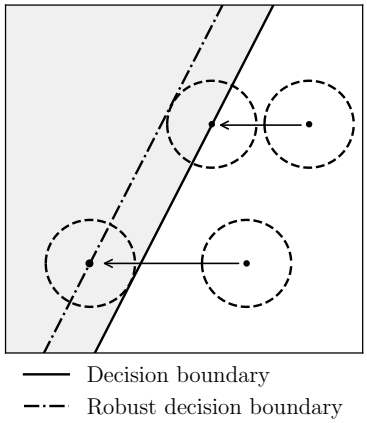

Figure 8: Standard recourse with respect to the robust decision boundary results in adversarially robust recourse.

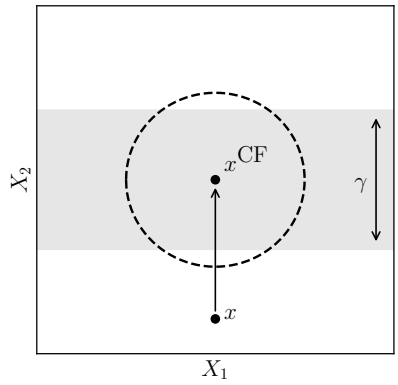

Figure 9: Classifier of Example 1. The shaded area is the favourably classified region of the feature space

C. Proofs

C.1. Theorem 1

Let $a^*(x) = do(\mathbf{X}_\mathcal{I} = x_\mathcal{I} + \theta^*)$ be a minimum-cost recourse action for some classifier h and individual $x \in \mathcal{X}$. Assume that a^* is a robust recourse action for the uncertainty set $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$. Let us choose some intervened-upon feature $\mathbf{X}_{\mathcal{I}_j}$ that is not a causal ancestor of any other intervened-upon feature $\mathbf{X}_{\mathcal{I}_i}$ $\forall i \neq k$. For a DAG causal graph, at least one such feature $\mathbf{X}_{\mathcal{I}_j}$ must clearly exist. We will consider perturbations along precisely this feature $\mathbf{X}_{\mathcal{I}_j}$.

Let $e^j \in \mathbb{R}^{|\mathcal{I}|}$ be a standard basis vector such that $e^j_j = 1$ and $e^j_i = 0$ $\forall i \neq j$. Consider the perturbation $\delta = \alpha e^j \operatorname{sign}(\theta_j)$ for any $\alpha \leq \epsilon$. The modified action $a'(x) = do(X_\mathcal{I} = x_\mathcal{I} + \theta^* - \delta)$ is a valid recourse action, since

$$\begin{aligned} h(\mathbb{CF}(x, a')) &= h(\mathbb{CF}(x, do(X_\mathcal{I} = x_\mathcal{I} + \theta^* - \delta)) \\ &= h(\mathbb{CF}(\mathbb{CF}(x, -\delta), do(X_\mathcal{I} = x_\mathcal{I} + \theta^*))) \\ &= h(\mathbb{CF}(\mathbb{CF}(x, -\delta), a^*)) \\ &= 1 \end{aligned} \tag{9}$$

where the second equality holds given that the SCM $\mathcal{M}$ is an additive noise model and $\mathbf{X}_j$ is not a causal ancestor of the intervened-upon features $\mathbf{X}_\mathcal{I}$. The last equality holds per the assumption that a^* is an adversarially robust recourse action, and since $\|\delta\| \leq \epsilon$. If the cost function is subadditive, then it must be that

$$\begin{aligned} c(x, a') &= c(x, \theta^* + \theta - \delta) \\ &< c(x, \theta^* + \theta) \\ &= c(x, a^*) \end{aligned} \tag{10}$$

Thus, a' is a valid recourse action which has strictly lower cost than a^*. This is a contradiction on a^* being a minimum-cost recourse action, which stems from the assumption that a^* is adversarially robust. Consequently, the minimum-recourse action a^* is provably fragile to arbitrarily small (i.e., $\epsilon > 0$) perturbations to the features of the individual x seeking recourse.

C.2. Proposition 1

Example 1. *Consider $\mathcal{X} = \mathbb{R}^2$ and $h(x) = \sin(2\gamma\pi^{-1}x_2) \geq 0$ for $0 < \gamma < \epsilon$. Assume that all features are independently modifiable, and consider the uncertainty set $B(x) = \{x + \Delta \mid \|\Delta\|_2 \leq \epsilon\}$. Whilst there exists some recourse recommendation for all $x \in \mathcal{X}$, there does not exist any adversarially robust recourse action for any $x \in \mathcal{X}$. See Figure 9.*

C.3. Proposition 2

Let h be a classifier for which there exists some robustly classified individual $x^+ \in \mathcal{X}$ such that $h(x^+) = 1$ $\forall x' \in B(x^+)$, where $B(x^+) = \{\mathbb{CF}(x^+, \Delta) | \, \|\Delta\| \leq \epsilon\}$. For any given individual x, the action $a(x) = do\left(\mathbf{X} = x + (x^+ - x)\right)$ results in

the counterfactual individual $\mathbb{CF}(x, a) = x^+$. The action a is a recourse action, since $h(x^{\text{CF}}) = h(x^+) = 1$. Moreover, the action a is feasible, since all features are actionable by assumption. For any $x' \in B(x)$ it holds that

$$\{\mathbb{CF}(x', a) \mid x' \in B(x)\} = \{x' + x^+ - x \mid x' \in B(x)\} = B(x^+) \tag{11}$$

where the last equality holds since the SCM $\mathcal{M}$ is linear. Consequently, it holds that

$$h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \; \forall x' \in B(x) \iff h(x') = 1 \; \forall x' \in B(x^+) \tag{12}$$

since the equality in the right hand side holds by assumption that x^+ is robustly classified, it follows that the action a is an adversarially robust recourse action.

C.4. Proposition 3

By assumption the classifier $h(x) = \langle w, x\rangle \geq b$ and SCM $\mathcal{M}$ are linear, and thus per Proposition 4 a recourse action a is adversarially robust if it holds that $\langle w, \mathbb{CF}(x, a)\rangle \geq b'$ for some $b' > b$. By assumption, there exists some feature $\mathbf{X}_j$ such that $\mathbf{X}_j$ is actionable and unbounded. Consider the recourse action $a(x) = do(\mathbf{X}_j = x_j + \theta)$ for $\theta \in \mathbb{R}$. Due to the linearity assumptions on the SCM, $\mathbb{CF}(x, a) = x + \theta v$ for some $v \in \mathbb{R}^n$. Then, $\langle w, \mathbb{CF}(x, a)\rangle = \langle w, x + \theta v\rangle = \langle w, x\rangle + \theta\langle w, v\rangle$. A robust recourse action is equivalent to any θ such that $\theta\langle w, v\rangle \geq b' - \langle w, x\rangle$. If $\langle w, v\rangle \neq 0$, then clearly it is possible to set θ to have arbitrarily large magnitude and same sign as $\langle w, v\rangle$, such that the inequality above is met. Since $\mathbf{X}_j$ is actionable and unbounded, $a(x) = do(\mathbf{X}_j = x_j + \theta)$ is a feasible action. Consequently, a is a robust recourse action.

We argue that the requirement $\langle w, v\rangle \neq 0$ is a mild condition on the weights of the classifier, precisely the non-trivial case where the weights of the classifier are not chosen adversarially to the SCM. Such condition is the causal counterpart of the trivial requirement that $w_j \neq 0$ in the statements presented by Ustun et al. (2019) for the non-causal recourse setting.

C.5. Proposition 4

The adversarially robust recourse problem is defined as

$$\min_{a(x)=do(X_{\mathcal{I}}=x_{\mathcal{I}}+\theta)} c(x, a) \quad \text{s.t.} \quad a \in \mathcal{F}(x) \; \wedge \; h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \quad \forall \, x' \in B(x) \tag{13}$$

For a linear classifier $h(x) = \langle w, x\rangle \geq b$ the uncertain constraint is equivalent to

$$h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \quad \forall \, x' \in B(x) \iff \left(\min_{x' \in B(x)} \langle w, \mathbb{CF}\left(x', a\right)\rangle\right) \geq b \tag{14}$$

Under the assumption that the SCM $\mathcal{M}$ is linear, for any plausible individual $x' = \mathbb{CF}(x, \Delta)$ it holds that

$$\begin{aligned}
\mathbb{CF}(x', a) &= \mathbb{S}^a\left(\mathbb{S}^{-1}\left(x'\right)\right) \\
&= \mathbb{S}^a\left(\mathbb{S}^{-1}\left(\mathbb{CF}(x, \Delta)\right)\right) \\
&= \mathbb{S}^a\left(\mathbb{S}^{-1}\left(\mathbb{S}^{\Delta}\left(\mathbb{S}^{-1}(x)\right)\right)\right) \\
&= \mathbb{S}^a\left(\mathbb{S}^{-1}\left(\mathbb{S}\left(\mathbb{S}^{-1}(x) + \Delta\right)\right)\right) \\
&= \mathbb{S}^a\left(\mathbb{S}^{-1}(x) + \Delta\right) \\
&= \mathbb{S}^a\left(\mathbb{S}^{-1}(x)\right) + \mathbb{S}^a\left(\Delta\right) \\
&= \mathbb{CF}(x, a) + J_{\mathbb{S}^{\mathcal{I}}}\Delta
\end{aligned} \tag{15}$$

where $J_{\mathbb{S}^{\mathcal{I}}}$ denotes the Jacobian of the interventional mapping $\mathbb{S}^{\mathcal{I}}$ resulting from hard-intervening on the variables $\mathbf{X}_{\mathcal{I}}$.

Then for the uncertainty set $B(x) = \{\mathbb{CF}(x, a) \mid \|\Delta\| \leq \epsilon\}$ the uncertain constraint in Equation 14 is equivalent to

$$\begin{aligned}
\min_{x' \in B(x)} \langle w, \mathbb{CF}\left(x', a\right)\rangle &= \min_{\|\Delta\| \leq \epsilon} \langle w, \mathbb{CF}\left(x, a\right) + J_{\mathbb{S}^{\mathcal{I}}}\Delta\rangle \\
&= \langle w, \mathbb{CF}\left(x, a\right)\rangle + \min_{\|\Delta\| \leq \epsilon} \langle w, J_{\mathbb{S}^{\mathcal{I}}}\Delta\rangle \\
&= \langle w, \mathbb{CF}\left(x, a\right)\rangle - \left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^* \epsilon
\end{aligned} \tag{16}$$

Consequently, the adversarially robust recourse problem reduces to

$$\min_{a(x)=do(X_{\mathcal{I}}=x_{\mathcal{I}}+\theta)} c(x,a) \quad \text{s.t.} \quad a \in \mathcal{F}(x) \ \wedge \ \langle w, \mathbb{CF}(x,a)\rangle \geq b + \left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^* \epsilon \tag{17}$$

This is equivalent to the standard recourse problem for the classifier $h'(x) = \langle w, x\rangle \geq b + \left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^* \epsilon$. See Figure 8.

C.6. Proposition 5

For a classifier $h(x) = \tilde{h}(x) \geq b$ where $\tilde{h} : \mathcal{X} \to (0, 1]$ the uncertain constraint is equivalent to

$$\begin{aligned} h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \quad \forall\, x' \in B(x) &\iff \left(\min_{x' \in B(x)} \tilde{h}\left(\mathbb{CF}\left(x', a\right)\right)\right) \geq b \\ &\iff \left(\min_{x' \in B(x)} \log \tilde{h}\left(\mathbb{CF}\left(x', a\right)\right)\right) \geq \log b \\ &\iff \left(-\max_{x' \in B(x)} -\log \tilde{h}\left(\mathbb{CF}\left(x', a\right)\right)\right) \geq \log b \\ &\iff \log b + \max_{x' \in B(x)} \ell\left(\tilde{h}\left(\mathbb{CF}\left(x', a\right)\right), \mathbf{1}\right) \leq 0 \end{aligned} \tag{18}$$

The corresponding Lagrangian is then

$$L(a, \lambda) = c(x, a) + \lambda\left(\log b + \max_{x' \in B(x)} \ell\left(\tilde{h}\left(\mathbb{CF}\left(x', a\right)\right), \mathbf{1}\right)\right) \tag{19}$$

Therefore, adversarially robust recourse problem is equivalent (Boyd & Vandenberghe, 2004) to

$$\min_{a(x)=do(X_{\mathcal{I}}=x_{\mathcal{I}}+\theta)\in\mathcal{F}(x)} \max_{\lambda \geq 0} L(a, \lambda) \tag{20}$$

C.7. Proposition 6 and extension to the causal setting

Let the classifier $h(x) = \langle w, x\rangle \geq b$ and SCM $\mathcal{M}$ be linear. By assumption, the action $a(x) = do\left(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta\right)$ is a recourse action for h, and the features $\mathbf{X}_{\mathcal{I}}$ are actionable and unbounded. Consider a modified action of the form $a'(x) = do\left(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + (1+\beta)\theta\right)$ where the individual x is asked to intervene on the features $x_{\mathcal{I}}$ by an additional factor β. Per the assumption that $x_{\mathcal{I}}$ are unbounded, the modified action a' is actionable. Per Proposition 4, a sufficient condition for a' to be an adversarially robust recourse action against the uncertainty set $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$ is

$$\langle w, \mathbb{CF}\left(x, a'\right)\rangle \geq b + \left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^* \epsilon \tag{21}$$

Since the SCM $\mathcal{M}$ is assumed linear, it holds that $\mathbb{CF}(x, a') = x + J_{\mathbb{S}^I}(1+\beta\epsilon)\theta$ (first order Taylor expansion). Then,

$$\begin{aligned} \langle w, \mathbb{CF}\left(x, a'\right)\rangle &= \langle w, x + (1+\beta\epsilon) J_{\mathbb{S}^{\mathcal{I}}}\theta)\rangle \\ &= \langle w, x + J_{\mathbb{S}^I}\theta\rangle + \beta\epsilon\langle w, J_{\mathbb{S}^{\mathcal{I}}}\theta\rangle \\ &= \langle w, \mathbb{CF}(x, a)\rangle + \beta\epsilon\langle w, J_{\mathbb{S}^{\mathcal{I}}}\theta\rangle \\ &\geq b + \beta\epsilon\langle w, J_{\mathbb{S}^{\mathcal{I}}}\theta\rangle \end{aligned} \tag{22}$$

where the last inequality follows by assumption that a is a recourse action for h. Consequently, we seek a condition on β such that Equation 22 satisfies the inequality in Equation 21. In particular, it holds that

$$\beta = \frac{\left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^*}{\langle w, J_{\mathbb{S}^{\mathcal{I}}}\theta\rangle}\epsilon \implies \langle w, \mathbb{CF}\left(x, a'\right)\rangle \geq b + \left\|J_{\mathbb{S}^{\mathcal{I}}}^T w\right\|^* \epsilon \implies a' \text{ is adversarially robust recourse.} \tag{23}$$

Since a is a recourse action, it must hold that $\langle w, J_{\mathbb{S}^{\mathcal{I}}}\rangle > 0$. Consequently, $0 < \beta < \infty$, and thus the additional change β to $\mathbf{X}_{\mathcal{I}}$ required by the individual to robustify the recourse action a is finite.

Under the assumption that the cost function is subadditive, it follows that $c(x, a') \leq (1 + \beta)\ c(x, a)$. Rearranging, we obtain the following upper bound on the additional effort required by the individual to robustify the recourse action a:

$$\frac{c(x, a') - c(x, a)}{c(x, a)} \leq \beta, \quad \text{for} \quad \beta = \frac{\left\| J_{\mathbb{S}^{\mathcal{I}}}^T w \right\|^*}{\langle w, J_{\mathbb{S}^{\mathcal{I}}} \theta \rangle} \epsilon \tag{24}$$

Under the IMF assumption (i.e., in the non-causal setting), $J_{\mathbb{S}^{\mathcal{I}}}$ is the identity matrix I. Let $m_{\mathcal{A}} \in [0, 1]^n$ (resp. $m_{\tilde{\mathcal{A}}}$) be the mask vector for the set of actionable features $\mathcal{A}$ (resp. unactionable features $\tilde{\mathcal{A}}$). Trivially, $m_{\mathcal{A}} + m_{\tilde{\mathcal{A}}} = \mathbf{1}$. Then

$$\begin{aligned}
\frac{c(x, a') - c(x, a)}{c(x, a)} &\leq \frac{\|w\|^*}{\langle w, \theta \rangle} \epsilon \\
&= \frac{\|(m_{\mathcal{A}} + m_{\tilde{\mathcal{A}}}) \odot w\|^*}{\langle (m_{\mathcal{A}} + m_{\tilde{\mathcal{A}}}) \odot w, \theta \rangle} \epsilon \\
&\leq \frac{\|m_{\mathcal{A}} \odot w\|^* + \|m_{\tilde{\mathcal{A}}} \odot w\|^*}{\langle m_{\mathcal{A}} \odot w, \theta \rangle + \langle m_{\tilde{\mathcal{A}}} \odot w, \theta \rangle} \epsilon \\
&= \frac{\|m_{\mathcal{A}} \odot w\|^* + \|m_{\tilde{\mathcal{A}}} \odot w\|^*}{\langle m_{\mathcal{A}} \odot w, \theta \rangle} \epsilon
\end{aligned} \tag{25}$$

where the last equality follows from the fact that the prescribed feature change θ necessarily only affects actionable features, and thus $m_{\tilde{\mathcal{A}}} \odot \theta = \mathbf{0}$. We thus arrive at the upper bound presented in Proposition 6.

D. Relation between adversarial robustness and robustness of the cost of recourse

In this work we focus on generating adversarially robust recourse. That is, we aim to offer a single recourse action that remains valid for plausible individuals similar to the individual x seeking recourse. A related notion of robustness is the robustness of the cost of recourse: the extent to which similar individuals are offered different but similarly costly recourse recommendations (von Kügelgen et al., 2022; Slack et al., 2021; Artelt et al., 2021), that is:

$$\delta(x) = \sup_{x \in B(x)} \left| \left(\min_{a \in \mathcal{F}(x')} c(x', a) \text{ s.t. } h(\mathbb{CF}(x', a)) = 1 \right) - \left(\min_{a \in \mathcal{F}(x)} c(x, a) \text{ s.t. } h(\mathbb{CF}(x, a)) = 1 \right) \right| \tag{26}$$

First, we show that the robustness of the cost of recourse is necessary for the existence of adversarially robust recourse

Proposition 7. *Let $x \in \mathcal{X}$ be some individual for which there exists at least one recourse action, and let $B(x)$ be the uncertainty set. If $\delta(x) = \infty$, then there does not exist any adversarially robust recourse action for the individual x.*

Proof. By assumption there exists some recourse action for x, and therefore $\delta(x) = \infty$ implies that there exists some individual $x' \in B(x)$ for which there does not exist any recourse action. It follows that there does not exist any robust recourse action a^* for the individual x, since a^* necessarily cannot be a recourse action for x'. □

Additionally, under certain conditions, the cost of adversarially robust recourse for an individual x can be used to upper bound the robustness of the cost of recourse $\delta(x)$ around that individual x.

Proposition 8. *For some classifier h, individual $x \in \mathcal{X}$, and uncertainty set $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$, if*

(i) Similar individuals have equal actionability constraints $\mathcal{F}(x) = \mathcal{F}(x')\ \forall x' \in B(x)$.

(ii) The cost of any given recourse action a is the same for all similar individuals, that is, $c(x, a) = c(x', a)\ \forall x' \in B(x)$.

Then, if there exists some adversarially robust recourse action a^*, the robustness of the cost of recourse is upper bounded by the cost of the adversarially robust recourse action a^*, that is, $\delta(x) \leq c(x, a^*)$.

Proof. For notational simplicity, let $r(x) = \min_{a \in \mathcal{F}(x)} c(x, a)$ s.t. $h(\mathbb{CF}(x, a)) = 1$. Then

$$\begin{aligned}
\delta(x) &= \max \left(\max_{x' \in B(x)} r(x') - r(x), r(x) - \min_{x' \in B(x)} r(x') \right) \\
&\leq \max \left(c(x, a^*) - r(x), r(x) \right) \leq \max \left(c(x, a^*), r(x) \right) \leq c(x, a^*)
\end{aligned} \tag{27}$$

□

E. Implementation details for the experiments

E.1. Datasets

E.1.1. Features used

COMPAS (Larson et al., 2016): we use the features "age", "race", "sex" and "priors count", which are the variables in the causal graph of Nabi & Shpitser (2018). We consider "priors count" actionable. As actionability constraints, we assume that "priors count" is non-negative and can only be decreased.

Adult (Kohavi & Becker, 1996): we use the features "sex", "age", "native-country", "marital-status", "education-num", "hours-per-week", which are the variables in the causal graph of Nabi & Shpitser (2018). We consider "education-num" and "hours-per-week" actionable. We assume that "education-num" is bounded within $[1, 16]$ (i.e., 16 discrete categories, but for simplicity we assume the variable to be real-valued), and that "hours-per-week" is bounded within $[0, 100]$.

South German Credit: we use the version of the dataset introduced by Groemping (2019), which corrects some coding errors in the original of German Credit dataset from the UCI repository (Murphy, 1994). We use all 20 features, with target variable "credit risk". We assume "duration" and "amount" as actionable, since these are the only real-valued features (together with "age", which we assume unactionable). As actionability constraints, we require "duration" and "amount" to be positive.

Bail (Schmidt & Witte, 1988): we use all features except "file" (the original test-train split) and "time" (which directly gives away the recidivism). We use as target variable "recid" (recidivism). For consistency with Ross et al. (2021), we consider the features "school" (number of years of formal school completed) and "rule" (number of prison rule violations reported) as actionable. For the actionability constraints, we assume that "school" is bounded within $[1, 19]$ (we assume the variable to be real-valued) and can only be increased, and that "rule" is non-negative and can only be decreased.

Loan: we use the semi-synthetic SCM described in Karimi et al. (2020) Appendix E.1.2. We sample 1000 individuals. We assume "education", "income" and "savings" to be actionable. For the actionability constraints, we assume that all three actionable variables can only be increased. Additionally, we assume that "education" is bounded by above to the maximum education level observed in the training data.

E.1.2. Standard deviation of the features used

We standardize real-valued features, and define the uncertainty set $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\|_2 \leq \epsilon\}$ with respect to the standardized features. Therefore, any given feature can be perturbed by at least $\pm\ \epsilon \times$ the feature's standard deviation. To provide the reader with a better sense of the magnitude of uncertainty for each ϵ considered, we list the standard deviation of the real-valued features of each dataset.

COMPAS "Age": 11.7 years, "Number of convictions": 4.74 *Adult* "Age" 13.6 years, "Education number" 2.57, "Hours per week" 12.3 *German* "Age": 11.3 years, "Loan duration": 12.0 months, "Loan amount": 2822 Deutsche Mark. *Bail* "Age": 9.66 years, "Number of priors": 2.91 "Years of school": 2.46, "Rule violations": 2.41, "Time served': 22.1 months, "Follow-up period": 3.44 months. *Loan* "Age": 11.0 years. The rest of the features lack meaningful units.

For instance, for the COMPAS dataset, robustifying against $\epsilon = 0.1$ uncertainty is equivalent to robustifying against perturbations to the feature "age" of at least $\pm\ 0.1 \times 11.7 = 1.17$ years.

E.2. Training of the classifiers

We train the decision-making classifiers $h(x)$ using one of the following training objectives:

- Empirical risk minimization (ERM): $\min_\psi \mathbb{E}\left[\ell(h_\psi(x), y)\right]$
- ERM with actionable features (AF): $\min_\psi \mathbb{E}\left[\ell(h_\psi(m_\mathcal{A} \odot x), y)\right]$
- ALLR: $\min_\psi \mathbb{E}\left[\ell(h_\psi(x), y) + \mu_2 \left\|m_{\tilde{\mathcal{A}}} \odot \nabla_x h(x)\right\| + \mu_1 \max_{\|\delta\| \leq \epsilon} \left|h(x+\delta) - \langle \delta, \nabla_x h(x)\rangle - h(x)\right|\right]$
- Ross et al. (2021): $\min_\psi \mathbb{E}\left[\ell(h_\psi(x), y) + \mu \min_\delta \ell(h(x + m_\mathcal{A} \odot d), \mathbf{1})\right]$

We use the binary cross-entropy loss as the loss function ℓ. For each of the above training objectives, we train five different classifiers, each with a different random seed and thus a different 80%-20% train-test split. We use Adam (Kingma & Ba,

2015) as the optimizer with a learning rate of 10^{-3} and a batch size of 100. To determine a suitable number of training epochs for each dataset and training objective, we train for 500 epochs and select the number of training epochs which leads to the best predictive performance in terms of accuracy and Mathews Correlation Coefficient (MCC). The resulting number of training epochs used are presented in Table 1.

A tacit assumption in most of the algorithmic recourse literature is that the decision threshold for classifiers of the form $h(x) = \tilde{h}(x) \geq b$ for $\tilde{h} : \mathcal{X} \to (0, 1)$ is set to $b = 0.5$. In practical applications, however, the decision threshold b is most often selected based on the nature of the classification problem at hand. For instance, in loan application settings the banking institution has a particular interest in minimizing false positives, since individuals which are mistakenly deemed low risk are very likely to result in monetary losses for the bank. We make the simplifying choice of, after training the classifier $\tilde{h}(x)$, choosing the decision threshold $b \in (0, 1)$ which maximizes the classifier's MCC. A classifier's MCC is a good metric to describe the classifier's confusion matrix with a single number, where higher MCC are associated with better predictive performance (Powers, 2020).

For the ALLR and Ross et al. (2021) regularizers, we perform hyperparameter search over the hyperparameter μ. We choose the regularization strength μ which most facilitate the existence of robust recourse while maintaining sufficient predictive performance. For the regularizer of Ross et al., we search over $\mu \in \{0.01, 0.1, 0.8, 1.5\}$ and find $\mu = 0.8$ to work best across all datasets and model types. This aligns with the findings of Ross et al. (2021). For ALLR with NN classifiers, we heuristically find that $\mu_1 = 3.0$ works well across all datasets. We additionally perform hyperparameter search over $\mu_2 \in \{0.01, 0.1, 0.5, 3.0\}$. The best-performing hyperparameters μ_2 are presented in Table 2.

Table 1: Training epochs for each training method, dataset and model type

Training method	Dataset & model type									
	COMPAS		Adult		Loan		German		Bail	
	LR	NN	LR	NN	LR	NN	LR	NN	LR	NN
ERM & AF	100	10	30	30	20	100	500	20	200	50
Ross et al.	20	10	20	80	30	20	20	20	40	100
ALLR	10	20	20	80	20	30	40	20	20	500

Table 2: Hyperparameter μ_2 used for ALLR

Model type	Dataset				
	COMPAS	Adult	Loan	German	Bail
LR	0.1	0.1	0.1	0.1	0.1
NN	0.1	0.5	0.01	0.5	0.01

E.3. Algorithm 1

E.3.1. Projecting to the uncertainty set

For ϵ-neighborhoods $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$, one need only project to the ϵ-ball of the norm $\|\cdot\|$, since in general $\max_{x' \in B(x)} f(x) = \max_{\|\Delta\| \leq \epsilon} f(\mathbb{CF}(x, \Delta))$. For certain choices of norm $\|\cdot\|$ (e.g., the l_2 norm), projecting to the ϵ-ball can be efficiently done in closed form.

E.3.2. Hyperparameter tuning

We set $N_{\text{max}} = 100$ according to our computational budget. To solve the inner maximization problem, we additionally allow a maximum of 50 gradient steps, for a total of $100 \cdot 50 = 5000$ optimization steps. We tune λ heuristically. If λ is too large, few recourse actions are found (low weight given to crossing the decision boundary), whereas if λ is too small, recourse actions tend to be overtly costly (low weight given to finding low-cost recourse). We find $\lambda = 1.0$ to work well across all datasets. Additionally, we find $\gamma = 0.9$ to work well, as λ then decreases relatively slowly (favoring low-cost recourse being found) but λ is nonetheless close to 0 after N_{max} iterations (favoring crossing the decision boundary).

E.4. Metrics considered

- Magnitude of min. perturbation invalidating recourse: for some classifier h, individual x and recourse action a, we use the C&W adversarial attack (Carlini & Wagner, 2017) to search for the minimum additive intervention Δ_{adv} which invalidates a, that is, $\Delta_{\text{adv}} = \arg\min_{\Delta} \|\Delta\|_2$ s.t. $h(\mathbb{CF}(\mathbb{CF}(x, \Delta), a)) = 0$. We report its magnitude $\|\Delta_{\text{adv}}\|_2$. While we tested a variety of adversarial attacks, including the fast gradient sign method (Goodfellow et al., 2015), projected gradient descent (Madry et al., 2018) and DeepFool (Moosavi-Dezfooli et al., 2016), we found C&W to work best.
- % recourse found: for some classifier h and uncertainty ϵ, we sample 1000 negatively classified individuals from the test set and use our proposed methods to generate recourse actions which are robust against ϵ uncertainty. Out of those individuals for which a valid recourse action is found, we then use the C&W attack described above to generate adversarial perturbations Δ_{adv}. We report the rate of individuals for which $\|\Delta_{\text{adv}}\|_2 \geq \epsilon$ (i.e., for which some recourse action is found and we are not able to invalidate such action) out of the original 1000 individuals.
- Relative cost of recourse: we aim to assess whether classifiers trained using the proposed model regularizer offer (robust) recourse at a higher or lower cost compared to classifiers trained using the standard ERM approach. For any given model regularizer, we consider the individuals for which valid (robust) recourse is found for both the ERM classifier and the regularizer classifier. We then report the mean cost of recourse offered to those individuals under the regularized decision-making classifiers. To improve clarity of presentation, for each dataset we normalize the mean cost of recourse found such that the maximum mean cost of recourse reported is 1.

F. Ablation study for the ALLR regularizer

The proposed ALLR regularizer comprises two penalty terms, one which encourages the decision-making classifier to behave locally linearly, and a second one that penalizes locally relying on unactionable features. The two penalty terms are weighted by the hyperparameters μ_1 and μ_2, respectively. To verify that both penalty terms are necessary to facilitate the existence of robust recourse, we use the ALLR hyperparameters presented in Table 2 to train classifiers with two additional regularization approaches related to ALLR: one for which $\mu_1 = 0$ and a second one for which $\mu_2 = 0$ (i.e., only one of the two penalty terms is considered at a time). We only consider NN classifiers, since for linear classifiers the first penalty term (to behave locally linearly) is trivially always 0. We present results in Figure 10. We observe that both penalty terms are necessary to facilitate the existence of robust recourse.

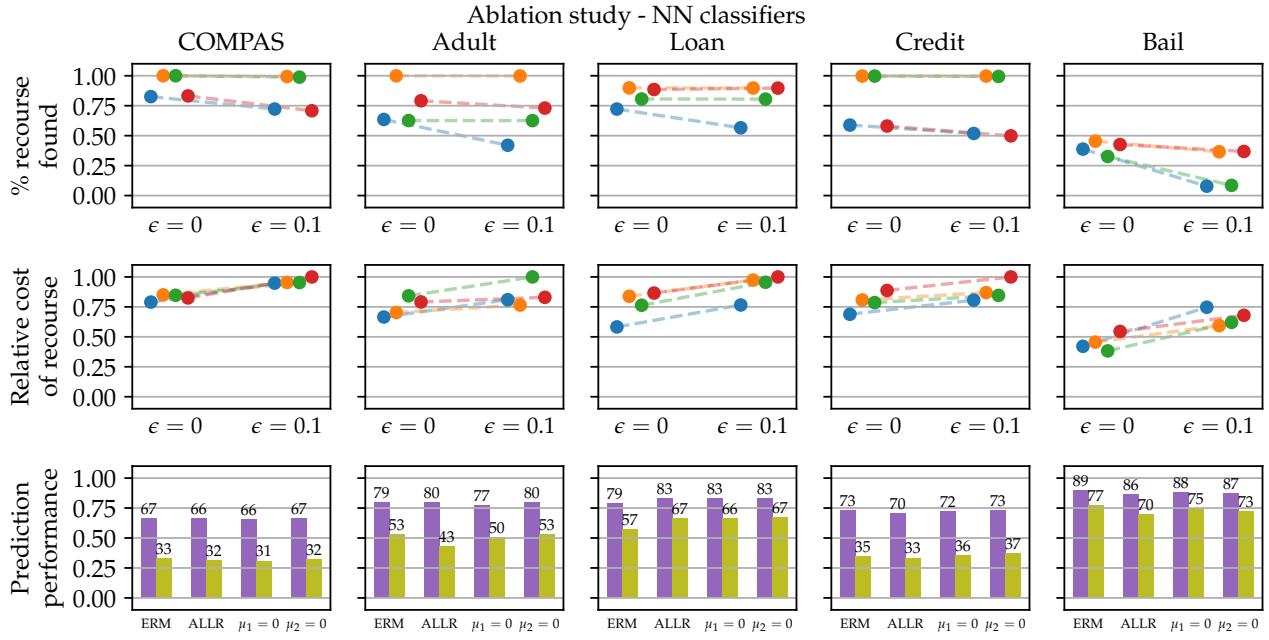

Figure 10: Results for the ablation study. Both penalties in ALLR are important to facilitate the existence of adversarially robust recourse. Legend: ERM ALLR ALLR $\mu_1 = 0$ ALLR $\mu_2 = 0$ Accuracy MCC score.