On the Adversarial Robustness of Causal Algorithmic Recourse

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

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

Abstract

Algorithmic recourse seeks to provide actionable recommendations for individuals to overcome unfavorable outcomes made by automated decision-making systems. The individual then exerts time and effort to positively change their circumstances. Recourse recommendations should ideally be robust to reasonably small changes in the circumstances of the individual seeking recourse. In this work, we formulate the adversarially robust recourse problem and show that methods that offer minimally costly recourse fail to be robust. We restrict ourselves to linear classifiers, and show that the adversarially robust recourse problem reduces to the standard recourse problem for some modified classifier with a shifted decision boundary. Finally, we derive bounds on the extra cost incurred by individuals seeking robust recourse, and discuss how to regulate this cost between the individual and the decision-maker.

1 Introduction

Machine learning (ML) classifiers are increasingly being used for consequential decision-making in domains such as justice and finance (e.g., granting pretrial bail or loan approval) [1]. The need to preserve human agency despite the rise in automatic decisions faced by individuals has motivated the study of algorithmic recourse, which aims to empower individuals by providing them with actionable recommendations to reverse unfavourable algorithmic decisions [2]. Prior works have argued that for recourse to warrant trust, the decision-maker must commit to reversing unfavourable decisions upon the decision-subjects fully adopting their prescribed recourse recommendations [3, 4, 5]. 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, such that the recommendations remain valid by the time the decision-subject is able to fully implement them.

For instance, consider a bank that commits to approving the loan of an individual if they increase their salary by some amount. However, by the time the individual achieves the prescribed salary increase the country's economic situation has slightly worsened and the classifier still deems the individual likely to default on the loan. Shielding the recourse recommendation 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. Therefore, we 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 [2]), or the presence of noise [6, 7], adversarial manipulation [8, 9] and other misrepresentations or errors [10]. Previous works on the robustness of recourse with respect to changes to the decision-subject have studied whether the *cost of recourse* is

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

robust [11, 12], that is, to what extent are similar individuals assigned recourse recommendations with similar cost. In contrast, we focus on the *validity of recourse*, and seek recourse recommendations which remain valid (i.e. lead to favourable classification outcomes) for *all* plausible changes to the features of the individual seeking recourse (e.g., all "similar" individuals). We argue that providing a recourse recommendation that remains valid over similar individuals is more desirable than providing multiple recommendations, with similar cost but potentially diverging actions, as the individual can more confidently commit to the valid set of actions.

We refer to this notion of robustness as the *adversarial robustness of recourse*, in order to distinguish it from other robustness considerations previously studied in the recourse literature (e.g., robustness of recourse with respect to changes to the decision-making classifier [13, 14, 15]). Indeed, by drawing inspiration from the adversarial robustness literature [8], we frame the adversarially robust recourse problem as a *minimax* optimization problem where the cost of recourse is minimized subject to adversarial perturbations to the features of the decision-subject seeking recourse.

We study the adversarial robustness of recourse from the lens of causality [16]. Causal recourse views recourse recommendations as causal interventions on the features of the decision-subject [17], and therefore presents a more faithful account of how the features of the individual change as the individual acts on its recourse recommendations, provided that the underlying structural causal model is known or can be approximated reasonably-well [18].

In this work, we consider the problem of robustifying recourse *ex-ante* against uncertainty in the features of the decision-subject. In Section 2, we discuss the different sources of uncertainty present in the recourse process and relate them with previous works on the robustness of recourse. In Section 3, we define a counterfactual notion of neighbourhoods of similar individuals. In Section 4, we formally define the adversarially robust recourse problem and we show that, in practice, minimum-cost recourse recommendations are fragile. In Section 5 we discuss how to generate adversarially robust recourse in the linear case. Finally, in Section 6 we discuss a theoretically motivated regularizer for training classifiers such that the extra cost of seeking robust recourse is reduced.

2 Background and related work

2.1 Structural causal models, interventions, and counterfactuals

We assume that the data-generating process of the features $\mathbf{X} = \{X_1, \ldots, X_n\}$ of individuals $x \in \mathcal{X}$ is characterised by a known *structural causal model* (SCM) [16] $\mathcal{M} = (\mathbf{S}, P_{\mathbf{U}})$. 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 $\mathbb{S}\left(\mathbb{S}^{-1}(x)\right) = x \quad \forall x \in \mathcal{X}$, 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* $do(\mathbf{X}_{\mathcal{I}} = \theta)$ [16] 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 $\mathbf{S}_i^{do(\mathbf{X}_{\mathcal{I}}=\theta)} = \mathbf{S}_i \;\; \forall i \notin \mathcal{I}$. Consequently, hard interventions sever the causal relationship between an intervened upon variables 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 [19]. In particular, *additive interventions* perturb the features $\mathbf{X}$ with some perturbation vector $\Delta \in \mathbb{R}^n$ while preserving all causal relationships. Additive interventions alter the structural equations according to $\mathbf{S}^{\Delta} = \left\{X_i := f_i\left(\mathbf{X}_{\text{pa}(i)}, \mathbf{U}_i\right) + \Delta_i\right\}_{i=1}^n$ [20].

Moreover, SCMs imply distributions over *counterfactuals*, allowing to reason about what would have happened under certain hypothetical interventions all else being equal. Under the afore-

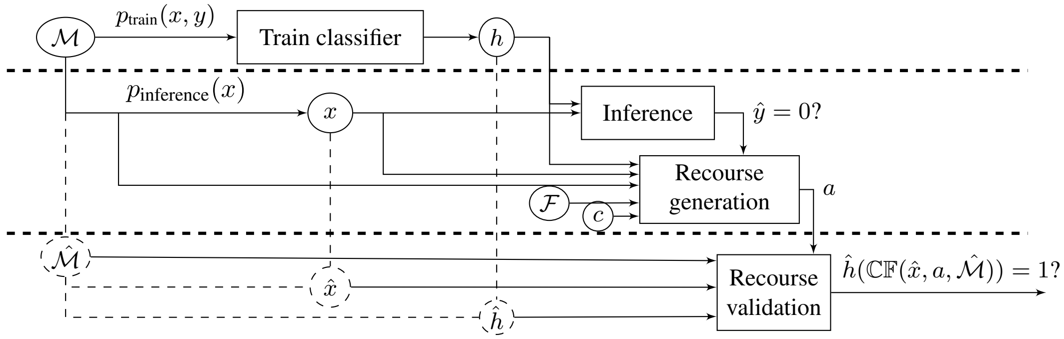


Figure 1: Overview of the recourse process. Elements where uncertainty may be present are represented with dashed circles. Possible relations between uncertain elements are represented with non-bold dashed lines. Bold dashed lines represent temporal jumps.

mentioned assumptions, 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 [16]. For notational convenience, we denote such mapping as $x^{\text{CF}} = \mathbb{CF}(x; do(\mathbf{X}_{\mathcal{I}} = \theta)) := \mathbb{S}^{do(\mathbf{X}_{\mathcal{I}}=\theta)}\left(\mathbb{S}^{-1}(x)\right)$ (resp. $x^{\text{CF}} = \mathbb{CF}(x; \Delta) := \mathbb{S}^{\Delta}\left(\mathbb{S}^{-1}(x)\right)$). We use the notation $x^{\text{CF}} = \mathbb{CF}(x; do(\mathbf{X}_{\mathcal{I}} = \theta), \mathcal{M})$ (resp. $x^{\text{CF}} = \mathbb{CF}(x; \Delta, \mathcal{M})$) to highlight that the counterfactual corresponds to a particular structural causal model $\mathcal{M}$.

2.2 The causal recourse problem

We consider the classification setting where some classifier $h : \mathcal{X} \to \{0, 1\}$ is used to assign either favourable or unfavourable outcomes to individuals $x \in \mathcal{X}$ (e.g., loan approval). Algorithmic recourse aims to provide unfavourably classified individuals a set of recommendations which if acted upon would result the individual being favourably classified [2, 5]. We adopt the causal view of recourse introduced by Karimi et al. [17] and model *recourse actions* as a hard interventions on the features of the individual seeking recourse, that is, $a = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$. We consider this additive form, rather than $a = do(\mathbf{X}_{\mathcal{I}} = \theta)$ as Karimi et al. [17], to explicitly allow the uncertainty in the factual individual x to propagate to the recourse action a.

For a recourse action a to be considered *valid*, the corresponding counterfactual individual must be favourably classified, that is, $h(\mathbb{CF}(x; a, \mathcal{M})) = 1$. Furthermore, 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 the individual x. Ideally, recourse recommendations should incur the least amount of effort possible for decision-subjects. The cost function $c(x, a)$ models the effort required by an individual $x \in \mathcal{X}$ to implement some recourse action a. Therefore, finding the minimum-cost recourse action for some individual $x \in \mathcal{X}$ is equivalent to solving the following optimization problem:

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

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

2.3 Uncertainties in the recourse process and robustness

Uncertainties may arise throughout the recourse process, as depicted in Figure 1, and may alter the validity of recourse, as highlighted in Equation 1. 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 in the form of perturbations, which may arise due to the presence of noise [6, 7], adversarial manipulation [8, 9] and other misrepresentations or errors in the data [10]. Regarding the classifier h, the optimization problem solved for model training often does not have unique optimal solution and multiple models may perform nearly equally well in the training data [21, 22]. 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. For instance, the distribution over inputs itself may change at inference time, under phenomena such as data-set shift [23, 24] or for tasks pertaining out of distribution generalization [25, 26].

From a causal perspective, changes in the distribution over inputs are a consequence of changes to the underlying SCM [27], which may also necessitate changes to the classifier. Indeed, the data-generation process characterised by the SCM $\mathcal{M}$ may be imperfectly known (as in most natural settings) or may dynamically change over time to some other SCM $\hat{\mathcal{M}} \in \mathcal{U}_{\mathcal{M}}$, where $\mathcal{U}_{\mathcal{M}}$ is the uncertainty set over future SCMs. 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 $\hat{h} \in \mathcal{U}_h$. Finally, it may be unreasonable to expect the individual x to not suffer changes outside of its control over a extended period of time (e.g., suffering an accident causing a decrease in savings), leading to uncertainty in the future individual $\hat{x} \in \mathcal{U}_x$. Thus, acting on the prescribed recourse may not lead to favourable classification due to changes to the SCM $\hat{\mathcal{M}}$, classifier $\hat{h}$, and/or factual individual $\hat{x}$.

2.4 Related work

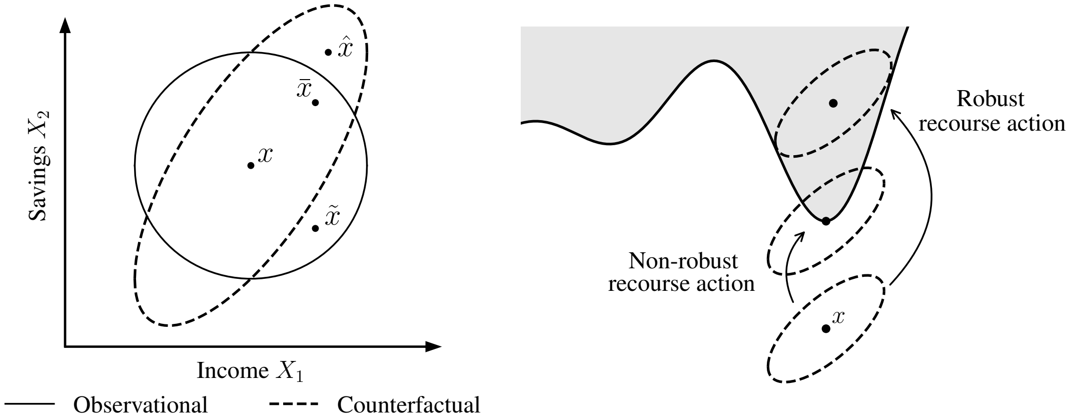
We now draw connections with existing literature on the robustness of recourse. Previous works have identified that perturbing the features of the decision-subject x may result in recourse recommendations with very different cost. Slack et al. [12] show that for gradient-based recourse methods it is possible to maliciously train a classifier such that small perturbations to the features of an individual drastically alter its cost of recourse, and von Kügelgen et al. [11] show that "counterfactual twins" obtained by intervening on sensitive attributes (e.g., race, gender) may be assigned recommendations with very different cost of recourse. While these works focus on the robustness of the cost of recourse with respect to changes to the individual, we instead focus on finding a *single* recourse action which remains valid under all plausible changes to the individual seeking recourse.

Other works have considered the problem of generating recourse actions which remain valid under changes to the classifier h. Pawelczyk et al. [28] show that recourse actions which place the counterfactual individual in regions of the feature space with large data support are more robust under predictive multiplicity. However, such recourse actions may be unnecessarily costly for the individual. In contrast, our approach seeks counterfactual individuals which are sufficiently far from the decision-boundary to be robust but not overtly so, thus ensuring minimum-cost robust recourse. Another line of work has considered the robustness of recourse with respect to changes to the classifier in response of dataset shift. Rawal et al. [29] show that recourse actions are typically not robust to such model changes, and Upadhyay et al. [14] aim to mitigate this issue through a *minimax* optimization procedure where the cost the recourse action is minimized subject to the action being valid under adversarial perturbations to the classifier h. In contrast, we focus on generating recourse under adversarial perturbations to the factual individual x, and we consider causal recourse.

Finally, Karimi et al. [18] consider the setting where the underlying SCM is not know and thus must be approximated, and propose a recourse 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. [18] and both approaches could be used in tandem.

3 Counterfactual similarity: a causal view on perturbations

In the adversarial robustness literature, the similarity of two data samples is often measured with some norm-induced metric $d(x, y) = \|x - y\|$. Then, the neighbourhood $B(x)$ of individuals similar to some individual x can be obtained by applying perturbations to x, that is, $B(x) = \{x + \delta | \|\delta\| \leq \epsilon\}$. Intuitively, small perturbations to the features of some individual x should result in individuals similar to x. For example, ∞-norm perturbations to the pixel values of an image result in new images which are indistinguishable from the original by the human eye [9].



(a) (b)

Figure 2: (a) Illustration of observational and counterfactual uncertainty sets $B(x)$ corresponding to 2-norm perturbations for the SCM $X_1 = U_1$, $X_2 = X_1 + U_2$. (b) Robust recourse actions ensure that every individual $x' \in B(x)$ in the uncertainty set is favourably classified.

From a causal perspective, these perturbations are equivalent to additive interventions on the features of the individual under the IMF assumption, that is, if not causal relations exist between features. However, if we have knowledge of data-generation process of the features x in the form of some structural causal model $\mathcal{M}$, then explicitly considering the causal relationships implied by $\mathcal{M}$ can potentially provide more informative neighbourhoods of individuals. In particular, perturbing some feature may have downstream causal effects on other features.

For instance, consider the linear SCM $\mathcal{M}$ with variables $X_1 = U_1$, $X_2 = X_1 + U_2$ respectively denoting the income and savings of some individual x. Figure 2a illustrates the similarity neighbourhoods corresponding to 2-norm perturbations under the IMF assumption (observational) and the true SCM $\mathcal{M}$ (counterfactual). For the latter, perturbations increasing the income of the individual (e.g., some salary bonus) also increase its savings, as one would expect. The counterfactual neighbourhood implies that the individual x is more similar to some counterfactual individual $\bar{x}$ with higher income and higher savings than to some other individual $\tilde{x}$ with higher income but lower savings, since the latter is not well explained by the SCM $\mathcal{M}$ and thus 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). Therefore, we ague that counterfactual similarity neighbourhoods may be inherently more informative than observational neighbourhoods. Furthermore, there are individuals in the counterfactual similarity set which do not belong to the observational similarity set, and vice versa. Therefore, generating robust recourse with respect to the observational set may fail to protect against plausible changes to the individual (e.g., $\hat{x}$), while needlessly overprotecting against changes which are not plausible under the SCM $\mathcal{M}$ (e.g., $\tilde{x}$).

Definition 1 (Neighbourhood of counterfactually similar individual)**.** *For some similarity norm $\|\cdot\|$ and SCM $\mathcal{M}$, the ϵ-neighbourhood of counterfactually similar individuals to some individual x is defined as the set of counterfactual individuals under all possible ϵ-small additive interventions*

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

We propose to model perturbations as additive interventions, such that perturbations to some feature have compounding downstream effects on its causal descendents. While we solely consider counterfactuals resulting from additive interventions, we recognize that other interventions could be considered, such as hard interventions or interventions on the distribution over exogenous variables $P_{\mathbf{U}}$. In the particular case of additive noise models, additive interventions are equivalent to shifts in the exogenous variables $\mathbf{U}$. We leave possible extensions for future work.

Table 1: Sufficient conditions for the universal existence of robust recourse

Classifier h	Actionability contraints	SCM $\mathcal{M}$	Existance of recourse	Existance of robust recourse
$\exists\, x \in \mathcal{X}$ s.t. $h(x) = 1$	All features actionable	Any	Yes (Ustun et al. [2])	No (Example 1)
$\exists\, x \in \mathcal{X}$ s.t. $h(x') = 1$ $\forall x' \in B(x)$	All features actionable	Any	Yes (Ustun et al. [2])	Yes (Lemma 1)
Linear	$\exists\, \mathbf{X}_j$ actionable and unbounded	Linear	Yes (Lemma 2)	Yes (Lemma 2)
Any	All bounded, ≥ 1 immutable	Any	No (Ustun et al. [2])	No (Follows directly)

4 The adversarially robust recourse problem

We consider the problem of generating recourse recommendations which remain valid under plausible perturbations to features of the individual x seeking recourse. We adopt a robust optimization view, requiring recourse actions a to remain valid for all individuals in the uncertainty set $B(x)$, that is, $h\left(\mathbb{CF}\left(x', a\right)\right) = 1 \;\; \forall x' \in B(x)$. Due to the similarity of this robustness formulation with that adopted in the adversarial robustness literature [8], we denote the corresponding recourse problem as the adversarially robust recourse problem.

Definition 2 (Adversarially robust recourse problem)**.** *For some uncertainty set $B(x)$, the minimum-cost recourse action which remains valid for all plausible individuals $x' \in B(x)$ is given by*

$$\underset{a=do(X_{\mathcal{I}}=x_{\mathcal{I}}+\theta)}{\arg\min} \;\; \max_{x \in B(x)} c(x', a) \quad \textit{s.t.} \quad a \in \mathcal{F}(x') \;\wedge\; h\left(\mathbb{CF}(x', a)\right) = 1 \tag{3}$$

In this section, we first show that under mild conditions, minimum-cost recourse is fragile to arbitrarily small changes to the features of the individual seeking recourse. Then, we derive sufficient conditions for the universal existence of adversarially robust recourse, as summarized in Table 1.

4.1 Recourse is fragile under mild conditions

Intuitively, minimum-cost recourse actions place the counterfactual individual arbitrarily close to the decision boundary of the classifier, since pushing the counterfactual further away from the decision boundary would incur additional cost. Consequently, small perturbations to the factual individual may result in small perturbations to the counterfactual individual such that it is negatively classified, as illustrated in Figure 2b. Theorem 1 states mild conditions under which minimum-cost recourse actions are indeed fragile to arbitrarily small perturbations to the features of the individual.

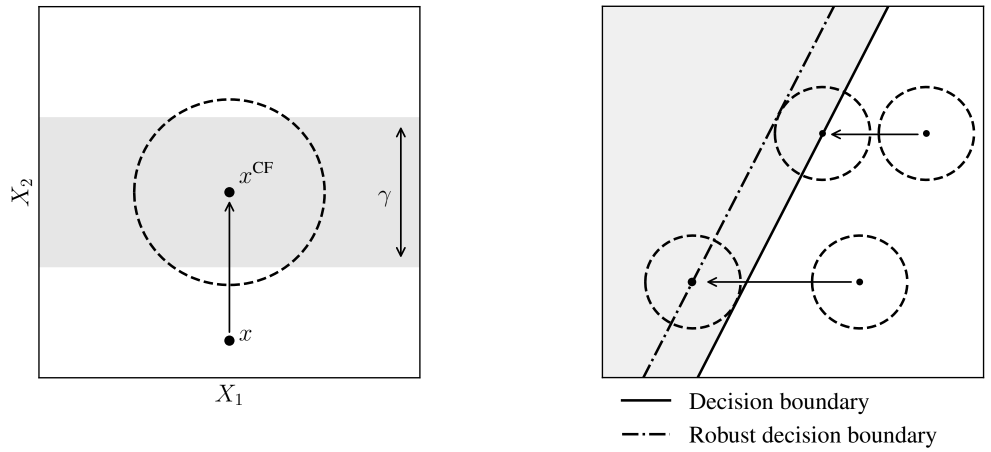
Theorem 1. *Let a^* be the solution to the recourse optimization problem stated in Equation 1. If*

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

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

Then for any $\epsilon > 0$ there exists $x' \in B(x) = \{\mathbb{CF}\left(x; \Delta)\right) \mid \|\Delta\| \leq \epsilon\}$ such that $h(\mathbb{CF}(x', a^*)) = 0$, that is, the recourse action a^* is invalidated for some arbitrarily small perturbation to x.

The conditions in Theorem 1 are often assumed by previous works. The first condition requires that larger changes to the features imply strictly more effort from the individual, which is satisfied by widely used cost functions such as weighted p-norms [18] or percentile costs [2]. The second condition 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, which is satisfied for the box actionability constrains



(a)

(b)

Figure 3: (a) Illustration of Example 1. The shaded area is the favourably classified region of the feature space. While there exists recourse for every individual, there does not exist robust recourse for any individual. (b) For a linear classifier, and under certain linearity assumptions on the SCM, the decision boundary of the classifier can be shifted to obtain a "robust decision boundary".

commonly assumed in the recourse literature (e.g., features are unbounded, bounded or immutable [3]). Therefore, in the settings commonly considered by the recourse literature, minimum-cost recourse actions are guaranteed to be fragile to perturbations to the factual individual x.

4.2 Sufficient conditions for the universal existence of robust recourse

The conditions required for the existence of robust recourse are strictly more restrictive than those required for the existence of recourse. Example 1, illustrated in Figure 3 (a), shows that even under the strong assumption that all features are actionable and that there exists recourse for every individual $x \in \mathcal{X}$, robust recourse may not exist for any individual $x \in \mathcal{X}$.

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

The above example relies on the fact that the classifier does not produce robust predictions for any $x \in \mathcal{X}$, and therefore no counterfactual can remain valid (i.e. favourably classified) in the presence of perturbations. This hints to some relation between robustness of prediction and robustness of recourse. In particular, it is necessary for the classifier to be minimally robust, in the sense that there must exist at least one individual $x^+ \in \mathcal{X}$ such that $h(x^+) = 1$ is robustly classified.

Lemma 1. *If all features are actionable and there exists some $x \in \mathcal{X}$ such that $h(x') = 1$ for all $x' \in B(x)$, then there exists some adversarially robust recourse recommendation for all $x \in \mathcal{X}$.*

In order to relax the condition of all features being actionable, we restrict ourselves to linear classifiers and linear SCM. Then, the existence of at least one actionable and unbounded feature is sufficient to guarantee the universal existence of robust recourse. Intuitively, the decision-maker can require arbitrarily large changes to an actionable and unbounded feature such that the counterfactual is favourably classified (e.g., increase savings in a loan application setting).

Lemma 2. *For a linear classifier $h(x) = \langle w, x \rangle \geq b$ and an SCM $\mathcal{M}$ with linear structural equations, if there exists a feature $\mathbf{X}_j$ such that $\mathbf{X}_j$ is actionable and unbounded and $w_j \neq 0$, then there exists at least one adversarially robust recourse action for all $x \in \mathcal{X}$.*

If all features are bounded and there exists at least one immutable feature, then as per Ustun et al. [2] Remark 3, it is not possible to guarantee the universal existence of recourse even in the linear case, and therefore it is also not possible to guarantee the universal existence of adversarially robust recourse.

5 Solving the adversarially robust recourse problem for the linear case

We now discuss how to solve the *minimax* optimization problem introduced in Section 4, such that recourse recommendations are robust to perturbations to the individual seeking recourse. We restrict ourselves to the linear case and show that, under certain assumptions, generating adversarially robust recourse for some classifier $h(x) = \langle w, x \rangle \geq b$ is equivalent to generating minimum-cost recourse for a modified classifier $h'(x) = \langle w, x \rangle \geq b'$ whose decision boundary is shifted such that $b' \geq b$. Intuitively, the "acceptance thereshold" of the classifier is increased, such that under all possible perturbations to the factual individual the corresponding counterfactuals nonetheless remain above the original decision threshold, as illustrated in Figure 3 (b).

Theorem 2. *Let $h(x) = \langle w, x \rangle \geq b$ be a linear classifier, $\mathcal{M}$ an SCM with linear structural equations, and $B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$ the uncertainty set of plausible individuals. If the feasibility set is invariant to perturbations to x, that is, $\forall x' \in B(x): \mathcal{F}(x) = \mathcal{F}(x')$, then the adversarially robust recourse problem is equivalent to the following optimization problem*

$$\min_{a = do(X_{\mathcal{I}} = x_{\mathcal{I}} + \theta)} c(x, a) \quad s.t. \quad a \in \mathcal{F}(x) \ \wedge \ \langle w, \mathbb{CF}(x, a) \rangle \geq b + \left\| J_{\mathbb{S}^{\mathcal{I}}}^{T} w \right\|^{*} \epsilon \tag{4}$$

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 features $\mathbf{X}_{\mathcal{I}}$.

Corollary 1. *Under the conditions of Theorem 2 and additionally under the IMF assumption, the adversarially robust recourse problem for the classifier $h(x) = \langle w, x \rangle \geq b$ is equivalent to the standard recourse problem for the modified classifier $h'(x) = \langle w, x \rangle \geq b + \|w\|^{*} \epsilon$.*

We highlight the importance of this result: if the conditions for Theorem 2 are satisfied, then generating recourse recommendations with respect to the modified classifier h' guarantees that recourse is adversarially robust. Conveniently, one may use any given recourse generation method from the rich corpus on algorithmic recourse in order to enforce, together with adversarial robustness, other desiderata such as large data-support [30, 28] or fairness constrains [31, 11].

6 A model regularizer to reduce the additional cost of recourse

In the previous sections, we have assumed a fixed classifier for which robust recourse must be generated. Then, to ensure that recourse recommendations are robust, individuals are asked to make more effort than they would have otherwise had to. Consequently, the burden of immunizing recourse against uncertainty falls solely on the decision-subjects. We argue, however, that robust recourse desiderata could be directly embedded into the training of the classifier. Satisfying such desiderata may come at a cost in predictive accuracy, thus shifting part of the burden of robust recourse from the decision-subject to the decision maker. In this section, we theoretically motivate a regularization penalty to reduce the additional cost of robust recourse.

We restrict ourselves to the linear case in order to derive an upper bound on the additional cost of robust recourse under certain actionability assumptions. For some recourse action $a = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$, we provide an expression for the increase in action magnitude β required for the more effortful recourse action $a' = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + (1 + \beta\epsilon)\theta)$ to be a robust recourse action.

Theorem 3. *Let h be a linear classifier $h(x) = \langle w, x \rangle \geq b$, $\mathcal{M}$ an SCM with linear structural equations, and $x \in \mathcal{X}$ a negatively classified individual for which there exists some recourse action $a = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + \theta)$. Then, there exists some constant*

$$\beta = \frac{\left\| J_{\mathbb{S}^{\mathcal{I}}}^{T} w \right\|^{*}}{\langle J_{\mathbb{S}^{\mathcal{I}}}^{T} w, \theta \rangle} \tag{5}$$

such that if $a' = do(\mathbf{X}_{\mathcal{I}} = x_{\mathcal{I}} + (1 + \beta\epsilon)\theta)$ is a feasible action $a' \in \mathcal{F}(x)$, then a' is an adversarial robust recommendation, that is, $\mathbb{CF}(x, a') = 1$ for all $x \in B(x) = \{\mathbb{CF}(x, \Delta) \mid \|\Delta\| \leq \epsilon\}$.

Corollary 2. *Under the assumptions of Theorem 3 and additionally under the assumption that the cost function is subadditive, then the additional cost incurred by robustifying action a is*

$$\frac{c(x, a') - c(x, a)}{c(x, a)} \leq \beta\epsilon \tag{6}$$

Consequently, $\beta\epsilon$ constitutes an upper bound on the additional cost of recourse incurred by the decision-subject as a result of seeking robust recourse. Observe that β (Equation 5) depends on the weights of the classifier w, and thus it may be possible to regularize w such that the upper bound on the additional cost of recourse $\beta\epsilon$ is reduced. For simplicity, we henceforth make the IMF assumption, such that $J_{\mathbb{S}^{\mathcal{I}}}^{T}$ is the identity matrix. Let $\mathcal{A}$ (resp. $\mathcal{U}$) be the set of features which are actionable (resp. unactionable) and $m_{\mathcal{A}} \in [0,1]^n$ (resp. $m_{\mathcal{U}} \in [0,1]^n$) the mask vector such that $(m_{\mathcal{A}})_i = 1 \iff i \in \mathcal{A}$ (resp. $(m_{\mathcal{U}})_i = 1 \iff i \in \mathcal{U}$). It is then possible decompose β into the weights corresponding to actionable features and those corresponding to unactionable features:

$$\beta = \frac{\|w\|^*}{\langle w, \theta \rangle} = \frac{\|m_{\mathcal{A}} \odot w + m_{\mathcal{U}} \odot w\|^*}{\langle w, m_{\mathcal{A}} \odot \theta \rangle} = \frac{\|m_{\mathcal{A}} \odot w\|^* + \|m_{\mathcal{U}} \odot w\|^*}{\langle m_{\mathcal{A}} \odot w, \theta \rangle} \tag{7}$$

where $\odot$ denotes the elementwise product. Consequently, reducing the dual norm $\|m_{\mathcal{U}} \odot w\|^*$ of the classifier weights corresponding to the unactionable features directly reduces the value of β, that is, the upper bound on the additional cost of robust recourse. The regularization term $\mu \|m_{\mathcal{U}} \odot w\|^*$ can be directly added to the loss function used for the training the classifier, where $\mu \in \mathbb{R}$ is the strength of regularization, inducing the learning bias "the classifier should rely more strongly on actionable features for its predictions". In subsequent versions of this manuscript, we plan to empirically evaluate the merits of this approach in reducing the additional cost of robust recourse.

7 Discussion

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 adopt an *ex-anti* approach to robustness of recourse by requiring the recourse recommendations to remain valid under adversarial perturbations to the features of the individual seeking recourse. We show that, in practice, minimum-cost recourse is fragile to arbitrarily small changes to the features of the individual. We then formulate the adversarially robust recourse problem, and show that generating robust recourse for a linear classifier h is equivalent to generating recourse for some modified linear classifier h' with a shifted decision boundary. Finally, we motivate a regularizer for training classifiers such that the additional cost of robust recourse is reduced, in order to regulate the bourden of robustness between the decision-maker and the decision-subject.

The approach to robustness of recourse presented in this work relies on the strong assumption that it is possible to model all relevant uncertainty in the recourse process. However, this task can be a very challenging in and of itself, if not unfeasible. Problematically, if there is uncertainty unaccounted for then there is a risk that the suggested recourse may not be robust enough. On the other hand, if the uncertainty sets are too extensive, then the recourse offered may be needlessly costly to implement. Furthermore, the conditions for the existence of robust recourse are strictly more restrictive than those of standard recourse, presenting an additional challenge over the the standard recourse problem.

In future versions of this manuscript, we plan to consider the adversarially robust recourse problem for non-linear classifiers, and experimentally validate the theoretical findings presented in this work.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org. *URL: http://www. fairmlbook. org*, 2019.

[2] 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.

[3] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[4] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 284–293, 2020.

[5] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[6] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. *Advances in Neural Information Processing Systems*, 29:1632–1640, 2016.

[7] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[10] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

[11] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *ICML 2021 Workshop on Algorithmic Recourse*, 2021.

[12] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. 2021.

[13] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

[14] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. 2021.

[15] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109*, 2021.

[16] Judea Pearl. *Causality*. Cambridge university press, 2009.

[17] 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*, pages 353–362, 2021.

[18] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, pages 265–277, 2020.

[19] Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pages 322–331. Springer, 2004.

[20] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

[21] Leo Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

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

[23] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[24] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[25] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[26] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[27] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[28] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pages 809–818. PMLR, 2020.

[29] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Can i still trust you?: Understanding the impact of distribution shifts on algorithmic recourses. *arXiv preprint arXiv:2012.11788*, 2020.

[30] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[31] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.