On the Impact of Adversarially Robust Models on Algorithmic Recourse

Satyapriya Krishna
Harvard University
skrishna@g.harvard.edu

Chirag Agarwal
Harvard University, Adobe
chiragagarwall12@gmail.com

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

Abstract

The widespread deployment of machine learning models in various high-stakes settings has underscored the need for ensuring that individuals who are adversely impacted by model predictions are provided with a means for recourse. To this end, several algorithms have been proposed in recent literature to generate recourses. Recent research has also demonstrated that the recourses generated by these algorithms often correspond to adversarial examples. This key finding emphasizes the need for a deeper understanding of the impact of adversarially robust models (which are designed to guard against adversarial examples) on algorithmic recourse. In this work, we make one of the first attempts at studying the impact of adversarially robust models on algorithmic recourse. We theoretically and empirically analyze the cost (ease of implementation) and validity (probability of obtaining a positive model prediction) of the recourses output by state-of-the-art algorithms when the underlying models are adversarially robust. More specifically, we construct theoretical bounds on the differences between the cost and the validity of the recourses generated by various state-of-the-art algorithms when the underlying models are adversarially robust vs. non-robust. We also carry out extensive empirical analysis with multiple real-world datasets to not only validate our theoretical results, but also analyze the impact of varying degrees of model robustness on the cost and validity of the resulting recourses. Our theoretical and empirical analyses demonstrate that adversarially robust models significantly increase the cost and reduce the validity of the resulting recourses, thereby shedding light on the inherent trade-offs between achieving adversarial robustness in predictive models and providing easy-to-implement and reliable algorithmic recourse.

1 Introduction

As machine learning (ML) models are increasingly being deployed in high-stakes domains such as banking, healthcare, and criminal justice, it becomes critical to ensure that individuals who have been adversely impacted (e.g., loan denied) by the predictions of these models are provided with a means for recourse. To this end, several techniques have been proposed in recent literature to provide recourses to affected individuals by generating counterfactual explanations which highlight what features need to be changed and by how much in order to flip a model's prediction [49, 9, 42, 30, 21, 20, 47][1]. For

[1]The terms counterfactual explanations [49], contrastive explanations [19], and recourse [42] have often been used interchangeably in prior literature.

2022 Trustworthy and Socially Responsible Machine Learning (TSRML 2022) co-located with NeurIPS 2022.

instance, [49] proposed a gradient-based approach which returns the nearest counterfactual resulting in the desired prediction. [42] proposed an integer programming-based approach to obtain actionable recourses for linear classifiers. More recently, [21, 22] leveraged the causal structure of the underlying data for generating recourses [3, 27, 30].

Prior research has also theoretically and empirically analyzed the properties of the recourses generated by state-of-the-art algorithms. For instance, several recent works [33, 30, 10, 40] demonstrated that the recourses output by state-of-the-art algorithms are not robust to small perturbations to input instances, underlying model parameters, and to the recourses themselves. More recently, [28] demonstrated that the recourses output by state-of-the-art algorithms are very similar to adversarial examples. This finding is critical because there have been several efforts in the literature on adversarial ML [17, 24, 4] to build adversarially robust models that are not susceptible to adversarial examples. However, the impact of such models on the quality and the correctness of the recourses output by state-of-the-art algorithms remains unexplored. The aforementioned connections between adversarial examples and recourses underscore the need for a deeper investigation of the impact of adversarially robust models (which are designed to guard against adversarial examples) on algorithmic recourse. Such an investigation becomes particularly critical as the need for adversarial robustness of predictive models as well as the ability to obtain easy-to-implement and reliable recourses have often been touted as the cornerstones of trustworthy and safe ML both by prior research as well as recent regulations [48, 15]. However, there is no prior work that investigates the relationship and/or the trade-offs between these two critical pillars of trustworthy and safe ML.

In this work, we address the aforementioned gaps by making the first ever attempt at studying the impact of adversarially robust models on algorithmic recourse. We theoretically and empirically analyze the cost (ease of implementation) and validity (probability of obtaining a positive model prediction) of the recourses output by state-of-the-art algorithms when the underlying models are adversarially robust. More specifically, we construct theoretical bounds on the differences between the cost and the validity of the recourses generated by various state-of-the-art algorithms (e.g., gradient-based [49, 25] and manifold-based [31] methods) when the underlying models are adversarially robust vs. non-robust (See Section 3). To this end, we first derive theoretical bounds on the differences between the weights (parameters) of adversarially robust vs. non-robust models and then leverage these to bound the differences in the costs and validity of the recourses corresponding to these two sets of models.

We also carried out extensive empirical analysis with multiple real-world datasets from diverse domains. This analysis not only validated our theoretical bounds, but also unearthed several interesting insights pertaining to the relationship between adversarial robustness of predictive models and algorithmic recourse. More specifically, we found that the cost differences between the recourses corresponding to adversarially robust vs. non-robust models increase as the degree of robustness of the adversarially robust models increases. We also observed that the validity of recourses worsens as the degree of robustness of the underlying models increases. We further probed these insights by visualizing the resulting recourses in low dimensions using t-SNE plots, and found that the number of valid recourses around a given instance reduces as the degree of robustness of the underlying model increases.

This work lies in the intersection of Algorithmic Recourse methods and Adversarial Robustness. Please refer Appendix A for a detailed discussion of related works.

Algorithmic Recourse. Several approaches have been proposed in recent literature to provide recourses to affected individuals [9, 49, 42, 45, 30, 27, 18, 22, 8]. These approaches can be broadly categorized along the following dimensions [47]: *type of the underlying predictive model* (e.g., tree based vs. differentiable classifier), *type of access* they require to the underlying predictive model (e.g., black box vs. gradient access), whether they encourage *sparsity* in counterfactuals (i.e., only a small number of features should be changed), whether counterfactuals should lie on the *data manifold*, whether the underlying *causal relationships* should be accounted for when generating counterfactuals, and whether the output produced by the method should be *multiple diverse counterfactuals* or a single counterfactual. In addition, [34] also studied how to generate global, interpretable summaries of counterfactual explanations. Some recent works also demonstrated that the recourses output by state-of-the-art techniques might not be robust, i.e., small perturbations to the original instance [11, 38], the underlying model [41, 33], or the recourse [32] itself may render the previously prescribed recourse(s) invalid. These works also formulated and solved minimax optimization problems to find *robust* recourses to address the aforementioned challenges.

Adversarial Examples and Robustness. Prior works have shown that complex machine learning models, such as deep neural networks, are vulnerable to small changes in input [39]. This behavior of predictive models allows for generating adversarial examples (AEs) by adding in nitesimal changes to input targeted to achieve adversary-selected outcomes [14]. Prior works have proposed several techniques to generate AEs using varying degrees of access to the model, training data, and the training procedure [6]. While gradient-based methods [14, 24] return the smallest input perturbations which flip the label as adversarial examples, generative methods constrain the search for adversarial examples to the training data-manifold. Finally, some methods generate adversarial examples for non-differentiable and non-decomposable measures in complex domains such as speech recognition and image segmentation. Prior works have shown that Empirical Risk Minimization (ERM) does not yield models that are robust to adversarial examples [14, 24]. Hence, to reliably train adversarially robust models, [26] proposed the adversarial training objective which minimizes the worst-case loss within some ϵ-ball perturbation region around the input instances.

Intersections between Adversarial ML and Model Explanations. There has been a growing interest in studying the intersection of adversarial ML and model explainability [15]. Among all the existing works focusing on this intersection, two explorations are relevant to our work [29, 36]. [36] studied the interplay between adversarial robustness and post hoc explanations and [29] demonstrated that gradient-based feature attribution methods (e.g., vanilla gradients, gradient times input, integrated gradients, smoothgrad) may severely violate the primary assumption of attribution – features with higher attribution are more important for model prediction – in case of non-robust models. However, their results also demonstrate that such a violation does not occur when the underlying models are robust to ℓ_2 and ℓ_∞ input perturbations. More recently, [28] demonstrated that recourses generated by certain state-of-the-art methods are very similar to adversarial examples, and also argued that the methods proposed to output recourses and adversarial examples are designed with similar goals of changing the input minimally in order to achieve the desired outcome. While the aforementioned works explored the connections between adversarial ML and model explanations, none of these works focus on analyzing the impact of adversarially robust models on the recourses output by state-of-the-art algorithms.

2 Preliminaries

Notation. In this work, we denote a classifier $f : \mathcal{X} \to \mathcal{Y}$ mapping features $\mathbf{x} \in \mathcal{X}$ to labels $y \in \mathcal{Y}$, where $\mathbf{x}$ is a d-dimensional feature vector. We define a non-linear activation function $\sigma(\cdot)$ such that $f(\mathbf{x}) = \sigma(h(\mathbf{x}))$, where $h(\mathbf{x})$ is the logits. In addition, we represent the non-robust and adversarially robust models using $f_{NR}(\mathbf{x})$ and $f_R(\mathbf{x})$. Below we describe the methodological frameworks used for comparing recourses generated from non-robust and adversarially robust models.

Adversarially Robust models. Despite the superior performance of machine learning (ML) models, they are susceptible to adversarial examples (AEs), i.e., inputs generated by adding in nitesimal perturbations to the original samples targeted to change prediction label [10]. The standard approach to ameliorate this problem is training a model using adversarial training which minimizes the worst-case loss within some perturbation region (the perturbation model) [23]. In particular, for a classifier f, parameterized by weights $\mathbf{w}$, loss function $\ell(\cdot)$, and training data $\{\mathbf{x}_i, y_i\}_{i=1,2,\ldots,n} \in \mathcal{D}_{train}$, the optimization problem of minimizing the worst-case loss within ℓ_p norm perturbation with radius ϵ is:

$$\min_{\mathbf{w}} \frac{1}{|\mathcal{D}_{train}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{train}} \max_{\boldsymbol{\delta}\in\Delta_{p,\epsilon}} \ell(f_{\mathbf{w}}(\mathbf{x}+\boldsymbol{\delta}), y), \tag{1}$$

where $\mathcal{D}_{train}$ denotes the training dataset and $\Delta_{p,\epsilon} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_p \leq \epsilon\}$ is the ℓ_p ball with radius ϵ centered around sample $\mathbf{x}$.

Algorithmic Recourse. One of the ways in which recourse can be realized is by explaining to affected individuals what features in their pro le need to change and by how much in order to obtain a positive outcome. Counterfactual explanations which essentially capture the aforementioned information can therefore be used to provide recourse. The terms "counterfactual explanations" and "algorithmic recourse" have, in fact, become synonymous in recent literature [19, 43, 46]. More speci cally, algorithms that try to find algorithmic recourses do so by finding a counterfactual $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$ that is closest to the original instance $\mathbf{x}$ and change the model's prediction $f(\mathbf{x} + \boldsymbol{\delta})$ to the target label. Next,

we describe three methods we use to understand the implications of adversarially robust models on algorithmic recourses.

Score CounterFactual Explanations (SCFE).Given the classi er $f(x) = (h(x))$ and a distance function $d : R^d \quad R^d ! \quad R_+$, [49] de ne the problem of generating a recourse $x^0 = x +$ for sample x by minimizing the following objective:

$$\arg\min_{x^0} (h(x^0) \quad s)^2 + \quad d(x^0, x); \tag{2}$$

where s is the target score for x^0, is the regularization coef cient, and $d(\ )$ is the distance between sample x and its counterfactual counterpart x^0.

C-CHVAE. Given a Variational AutoEncoder (VAE) model with encoder E and decoder G trained on the original data distribution D_{train}, C-CHVAE [31] aims to generate recourses in the latent space Z, where $E : X ! Z$. The encoder E transforms a given sample x into a latent representation $z 2 Z$ and the decoder G takes z as input and produces x as similar as possible to x. To this end, given a sample x, C-CHVAE generates the recourse using the following objective function:

$$= \arg\min_{2Z} k \quad k \quad \text{such that } f(G(E(x) + \ )) \ 6= f(x); \tag{3}$$

where E allows to search for counterfactuals in the data manifold and projects the latent counterfactuals to the feature space.

Growing Spheres Method (GSM).While the above techniques directly optimize speci c objective functions for generating counterfactuals, GSM [25] uses a search-based algorithm to generate recourses by randomly sampling points around the original instance until a sample with the target label is found. In particular, GSM rst draws an $_2$-sphere around a given instance x, randomly samples point within that sphere, and checks whether any sampled points result in target prediction. Finally, they contract or expand the sphere until a (sparse) counterfactual is found and nally returned. GSM de nes a minimization problem using a function $c : X \quad X ! \quad R_+$, where $c(x; x^0)$ is the cost of moving from instance x to counterfactual x^0.

$$x^0 = \arg\min_{x^0 2 X} \{ c(x; x^0) \ j \ f(x^0) \ 6= f(x) \}; \tag{4}$$

where x^0 is sampled from the $_2$-ball around x such that $f(x^0) \ 6= f(x)$, $c(x; x^0) = kx^0 \quad x k_2 + kx^0 \quad x k_0$, and $2 R_+$ is the weight associated to the sparsity objective.

3 Theoretical Analysis

Next, we carry out a detailed theoretical analysis to bound the cost and validity differences of recourses generated by state-of-the-art recourse methods when the underlying models are adversarially robust vs. non-robust. More speci cally, we compare the cost differences w.r.t. the recourses obtained using 1) gradient-based methods such as SCFE [49] (Sec. 3.1.1) and 2) manifold-based methods such as C-CHVAE [31] (Sec. 3.1.2). Finally, we show that the validity of algorithmic recourse generated using existing methods for robust models is lower compared to that of non-robust models (Sec. 3.2).

3.1 Cost Analysis

The cost of a generated algorithmic recourse is de ned as the distance $`$ (e.g., $`_2$ distance) between the input instance x and the counterfactual x^0 obtained using a state-of-the-art recourse nding method [47]. Algorithmic recourses with lower costs are considered better since they enable minimal changes to input to achieve the desired outcome. Here, we theoretically analyze the cost difference of generating recourses using algorithmic recourse methods when the underlying models are non-robust and adversarially robust.

3.1.1 Gradient-based method: SCFE

Next, we carry out a detailed theoretical analysis to bound the cost and validity difference of recourses generated by state-of-the-art recourse methods when the underlying models are adversarially robust vs. non-robust.

Here, we derive the lower and upper bound for the cost difference of recourses generated by the SCFE [49] method when the underlying models are adversarially robust vs. non-robust. Following

previous works [13, 16, 29, 35, 43], we focus on locally linear model approximations as this lays the foundation for understanding non-linear model behavior. For the cost difference, we first define the closed-form solution for the optimal cost required to generate a recourse.

Definition 1. (Optimal Cost from [29]) For a given scoring function f with weights w the SCFE method generates a recourse x^0 for an input x using cost ℓ such that:

$$\ell = |m| \frac{}{+ \|w\|_2^2} \|w\|; \tag{5}$$

where $m = s - h(x^0)$ is the target residual, s is the target score for x, $h(x)$ is a local linear score approximation, and λ is a given hyperparameter.

Theorem 1. (Cost difference for SCFE) For a given instance x, let x^0_{NR} and x^0_R be the recourse generated using Wachter's algorithm for non-robust and adversarially robust models. Then, for a normalized Lipschitz activation function $\sigma(\cdot)$, the cost difference for the recourse generated for both models can be bounded as:

$$\left|\frac{|m_{NR}|}{\|w_{NR}\|_2} - \frac{|m_R|}{\|w_R\|_2}\right| \|x^0_{NR} - x^0_R\|_2 \leq \frac{}{+ \|w_{NR}\|^2} \|w_{NR}\| + \frac{}{+ \|w_R\|^2} \|w_R\|; \tag{6}$$

where w_{NR} and w_R are the weights of the non-robust and adversarially robust models, λ is the regularization coefficient in Wachter's algorithm, $m_{NR} = s - h_{NR}(x^0)$, $m_R = s - h_R(x^0)$ are the target residuals for robust ($f_R(x) = \sigma(h_R(x))$) and non-robust models ($f_{NR}(x) = \sigma(h_{NR}(x))$), respectively.

Proof Sketch. We derive the cost difference of recourses generated for non-robust and adversarially robust models by comparing their optimal solutions. Similar to [29], the upper bound results follow from Cauchy-Schwartz and triangle inequality. In addition, we also leverage reverse triangle inequality to derive a lower bound for the recourse difference. See Appendix B.1 for the complete proof.

The equality of Equation 6 entails that the upper bound of the recourse difference will have a tighter bound if the ℓ_2-norms of the weights w_R and w_B are bounded, and the lower bound of the recourse difference will be tighter if the output score of the non-robust and adversarially robust models is similar for the given sample x. □

3.1.2 Manifold-based method: C-CHVAE

We extend our analysis of bounding the cost difference of generated recourses using manifold-based methods for non-robust and adversarially robust models. In particular, we leverage C-CHVAE [31], that leverages variational autoencoders to generate counterfactuals. For a fair comparison, we assume that both models use the same encoder E and decoder G networks for learning the latent space of the given input space $\mathcal{X}$.

Definition 2. ([5]) An encoder model E is L-Lipschitz if $\forall z_1, z_2 \in \mathcal{Z}$, we have:

$$\|E(z_1) - E(z_2)\|_p \leq L\|z_1 - z_2\|_p. \tag{7}$$

Using Definition 7, we now derive the lower and upper bounds of the cost difference of recourses generated for non-robust and adversarially robust models.

Theorem 2. (Cost difference for C-CHVAE) Let z_{NR} and z_R be the generated recourse from C-CHVAE [31] method in the latent space using an L-Lipschitz generative model $G(\cdot)$ for a non-robust and adversarially robust model. Then, by definition of C-CHVAE, $x_{NR} = G(z_{NR}) = x + \epsilon_{NR}$ and $x_R = G(z_R) = x + \epsilon_R$ are the corresponding recourses in the input space. The cost difference between the recourses can then be bounded as:

$$L(r_R - r_{NR}) \leq \|\epsilon_R - \epsilon_{NR}\|_p \leq L(r_R + r_{NR}); \tag{8}$$

where L is the Lipschitz constant of the generative model, and r_{NR} and r_R be the corresponding radii chosen by the algorithm such that they successfully return a recourse for the non-robust and adversarially robust model.

Proof Sketch. The proof follows from Definition 7 and the triangle inequality. It shows that the cost difference for generating recourses using C-CHVAE is bounded by the product of the Lipschitz constant of the generative model and the radii chosen by the C-CHVAE to generate counterfactuals for the underlying non-robust and adversarially robust models. See Appendix B.2 for detailed proof. □

3.2 Validity Analysis

The validity of a given recourse x^0 is defined as the probability that it results in the desired outcome [47], denoted by $Pr(f(x^0) = 1)$. Below, we analyze the validity of the recourses by first deriving the upper bound of the difference in non-robust and adversarially robust model weights, and then use this lemma to show that the validity of non-robust model is higher than for the adversarially robust model.

Lemma 1. (Difference between non-robust and adversarially robust model weights) For a given instance x, let w_{NR} and w_R be weights of the non-robust and adversarially robust model. Then, for a normalized Lipschitz activation function $\sigma(\cdot)$, the difference in the weights w can be bounded as:

$$\|w_{NR} - w_R\|_2 \leq \eta \cdot (\gamma \|x\|_2 + \epsilon\sqrt{d}) \tag{9}$$

where η is the learning rate, ϵ is the ℓ_2-norm perturbation ball, y is the label for x, n is the total number of training epochs, and d is the dimension of the input features.

Proof Sketch. We derive the upper bound of the difference in non-robust and adversarially robust model weights, denoted by w, and show that it is proportional to the dimension of the input features times the ℓ_2 perturbation ball around the sample x. See Appendix B.3 for the detailed proof. □

Next, we show that the probability of a recourse action resulting in the desired outcome is greater for a non-robust model compared to that of the adversarially robust model.

Theorem 3. (Validity Comparison) For a given instance $x \in \mathbb{R}^d$ and desired target label denoted by unity, let x_R and x_{NR} be the counterfactuals for adversarially robust $f_R(x)$ and non-robust $f_{NR}(x)$ models respectively. Then, $Pr(f_{NR}(x_{NR}) = 1) \geq Pr(f_R(x_R) = 1)$ if $|f_{NR}(x_R) - f_{NR}(x_{NR})| \leq \eta \cdot (\gamma \|x\|_2 + \epsilon\sqrt{d}) \|x_R\|$, where η is the learning rate, ϵ is the ℓ_2-norm perturbation ball, y is the label for x, and n is the total number of training epochs.

Proof Sketch. We derive the difference between the probability that a valid recourse exists for a non-robust and adversarially robust model. Using data inequalities and Cauchy-Schwartz, we show that the condition for the validity is dependent on the weight difference of the models (Lemma 1). See Appendix B.4 for the detailed proof. □

4 Experimental Evaluation

In this section, we empirically analyze the impact of adversarially robust models on the cost and validity of recourses. First, we empirically validate our theoretical bounds on differences between the cost and validity of recourses output by state-of-the-art recourse generation algorithms when the underlying models are adversarially robust vs. non-robust. Second, we carry out further empirical analysis to assess the differences in cost and validity of the resulting recourses as the degree of the adversarial robustness of the underlying model changes on three real-world datasets.

4.1 Experimental Setup

Here, we describe the datasets used for our empirical analysis along with the predictive models, algorithmic recourse generation methods, and the evaluation metrics.

Datasets. We use three real-world datasets for our experiments: 1) The German Credit dataset[2] comprises demographic (age, gender), personal (marital status), and financial (income, credit duration) features from 1000 credit applicants, with each sample labeled as "good" or "bad" depending on their credit risk. The task is to successfully predict if a given individual is a "good" or "bad" customer in terms of associated credit risk. 2) The Adult dataset[3] contains demographic (e.g., age, race, and gender), education (degree), employment (occupation, hours-per week), personal (marital status, relationship), and financial (capital gain/loss) features for 48,842 individuals. The task is to predict if an individual's income exceeds \$50K per year. 3) The COMPAS dataset[4] has criminal records and demographics features for 18,876 defendants who got released on bail at the U.S state courts during

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[3] https://archive.ics.uci.edu/ml/datasets/Adult/

[4] https://github.com/propublica/compas-analysis

Figure 1: Empirically calculated cost differences (in orange) and our theoretical lower (in blue) and upper (in green) bounds for (a) C-CHVAE and (b) SCFE recourses corresponding to adversarially robust (trained using $\epsilon = 0.3$) vs. non-robust models trained on the Adult dataset. Figure (c) is the empirical difference between the validity of recourses for non-robust and adversarially robust model. Results show no violations of our theoretical bounds. See Appendix C for results using different values.

1990-2009. The dataset is designed to train a binary classifier to classify defendants into bail (i.e., unlikely to commit a violent crime if released) vs. no bail (i.e., likely to commit a violent crime).

Predictive models. We generate recourses for the non-robust and adversarially robust version of Logistic Regression (linear) and Neural Networks (non-linear) models. We use two linear layers with sigmoid activation functions as our predictor and set the number of nodes in the intermediate layers to twice the number of nodes in the input layer, which is the size of the input dimension in each dataset.

Algorithmic Recourse Methods. We analyze the cost and validity for non-robust and adversarially robust models w.r.t. three popular classes of recourse generation methods, namely, gradient-based (SCFE), manifold-based (C-CHVAE), and random search-based (GSM) methods (described in Sec. 2).

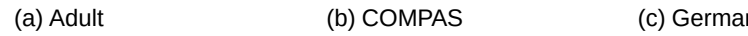
Evaluation metrics. To concretely measure the impact of adversarial robustness on algorithmic recourse, we analyze the difference between cost and validity metrics for recourses generated using non-robust and adversarially robust model. To quantify the cost, we measure the average cost incurred to act upon the prescribed recourses across all test-set instances, i.e., $\text{Cost}(x; x') = \frac{1}{|D_{test}|} \|x - x'\|_2$, where x is the input and x' is its corresponding recourse. To measure validity, we compute the probability of the generated recourse resulting in the desired outcome, i.e., $\text{Validity}(x; x') = \frac{|f : x' : f(x') = 1 \wedge x' = g(x; f)|}{|D_{test}|}$, where $g(x; f)$ returns recourses for input x and predictive model f.

Implementation details. We train non-robust and adversarially robust predictive models from two popular model classes (logistic regression and neural networks) for all three datasets. In the case of adversarially robust models, we adopt the commonly used min-max optimization objective for adversarial training using varying degree of robustness, i.e., $\epsilon \in \{0, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.3\}$. Note that the model trained with $\epsilon = 0$ is the non-robust model. Following [2] and [12], we pre-processed the input data by removing categorical features for efficient training of our models. We follow [29] to set the hyperparameters for the algorithmic recourse methods.

4.2 Empirical Analysis

Next, we describe the experiments that we carried out to understand the impact of adversarial robustness of predictive models on algorithmic recourse. More specifically, we will discuss (1) empirical verification of our theoretical bounds, (2) empirical analysis of the differences between the costs of recourses corresponding to non-robust vs. adversarially robust models, and (3) empirical analysis to compare the validity of the recourses corresponding to non-robust vs. adversarially robust models.

Empirical Verification of Theoretical Bounds. We empirically validate our theoretical findings from Section 3 on real-world datasets. In particular, we first estimate the empirical bounds (RHS of Theorems 1-2) for each instance in the test set by plugging the corresponding values of the parameters in the theorems and compare them with the empirical estimates of the cost differences between recourses generated using gradient- and manifold-based recourse methods (LHS of Theorems 1-2).



(a) Adult (b) COMPAS (c) German Credit

Figure 2: Analyzing cost differences between recourse generated using non-robust and adversarially robust neural networks for (a) Adult (b) COMPAS (c) German Credit datasets. We find that the cost difference (i.e., ℓ_2 norm) between the recourses generated for non-robust and adversarially robust models increases for increasing values of.

Figure 1 show the results obtained from the aforementioned analysis of cost differences. We observe that our bounds are tight, and the empirical estimates fall well within our theoretical bounds. Similarly, we observe that the validity of the non-robust model, as denoted by $Pr(y_{NR}(x) = 1)$ in Theorem 3, was higher than the validity of the adversarially robust model for all the test samples in Adult, German Credit, COMPAS datasets, following the condition in Theorem 3 for a large number of training iterations used for training adversarially robust models with $\in \{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

Cost Analysis. To analyze the impact of adversarial robustness on the cost of recourses, we compute the difference between the cost for obtaining a recourse using non-robust and adversarially robust model and plotted this difference for varying degrees of robustness. Results in Figure 2 show a significant increase in incurred costs to find algorithmic recourse for adversarially robust models compared to the non-robust model for all the datasets with increasing degrees of robustness. We observe a similar trend for the case of logistic regression, as shown in Figure 6 in Appendix C. Further, we observe a relatively smoother increasing trend for cost differences in the case of SCFE compared to others, which can be attributed to the stochasticity present in C-CHVAE and GSM. We also observe a higher cost difference in SCFE for most datasets, which could result from the larger sample size used in C-CHVAE and GSM. We observe a similar trend in cost differences when the sample size per iteration is reduced, which also resulted in more iterations to find recourse.

Validity Analysis. To analyze the impact of adversarial robustness on the validity of recourses, we compute the fraction of recourses resulting in the desired outcome, generated using non-robust and adversarially robust model under resource constraints, and plot it against varying degrees of robustness. Results in Figure 3 show that there is an even stronger impact of adversarial training on validity for neural networks trained on the three datasets. We observe a similar pattern for the case of the logistic regression model trained on the three datasets, shown in Appendix C. On average, we observe that the validity drops to zero for models adversarially trained with ≥ 0.2. To understand this further, we use t-SNE visualization [44] – a non-linear dimensionality reduction technique – to map points in the dataset to two-dimensional space and demonstrate a gradual decline in valid recourses around a local neighborhood with increasing in Figure 4, where x and y be the names of reduced dimensions. This decline suggests that a large number of recourses in the neighborhood of the input sample are now being classified with the same class as the input. Hence, this supports our hypothesis that adversarially robust models severely impact the validity of recourses and make the recourse search computationally expensive.

5 Conclusion

In this work, we theoretically and empirically analyzed the impact of adversarially robust models on algorithmic recourse. We theoretically bounded the differences between the costs of the recourses output by two state-of-the-art counterfactual explanation methods (SCFE and C-CHVAE) when

(a) Adult (b) COMPAS (c) German Credit

Figure 3: Analyzing validity of recourse generated using non-robust and adversarially robust neural networks for (a) Adult (b) COMPAS (c) German Credit datasets. We find that the validity decreases for increasing values of.

Figure 4: A t-SNE visualization of the change in availability of valid recourses (orange) for adversarially robust models trained using= [0; 0:15; 0:25], where a non-robust model is a model trained using = 0. Results are shown for a neural network model trained on the Adult dataset. We observe fewer valid recourses for higher values of in this local neighborhood.

the underlying models are adversarially robust vs. non-robust. In addition, we also bounded the differences between the validity of the recourses corresponding to adversarially robust and non-robust models. We empirically validated our theoretical results using three real-world datasets (Adult, COMPAS, and German Credit) and two popular model classes (neural networks and logistic regression). Our theoretical and empirical analyses demonstrated that adversarially robust models significantly increase the cost and reduce the validity of the resulting recourses, thereby highlighting the inherent trade-offs between achieving adversarial robustness in predictive models and providing easy-to-implement and reliable algorithmic recourses. Our work also paves the way for several interesting future research directions at the intersection of algorithmic recourse and adversarial robustness in predictive models. For instance, given the aforementioned trade-offs, it would be interesting to develop novel techniques which enable end users to navigate these trade-offs based on their personal preferences – e.g., an end user may choose to sacrifice the adversarial robustness of the underlying model in order to secure lower cost recourses.

6 Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful feedback and all the funding agencies listed below for supporting this work. This work is supported in part by the NSF awards IIS-2008461 and IIS-2040989, and research awards from Google, JP Morgan, Amazon, Harvard Data Science Initiative, and D3 Institute at Harvard. HL would like to thank Sujatha and Mohan Lakkaraju for their continued support and encouragement. The views expressed here are those of the authors and do not reflect the official policy or position of the funding agencies.

References

[1] C. Agarwal, A. Nguyen, and D. Schonfeld. Improving robustness to adversarial examples by encouraging discriminative features. In ICIP. IEEE, 2019.

[2] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, and M. Detyniecki. Imperceptible adversarial attacks on tabular data. arXiv preprint arXiv:1911.03274, 2019.

[3] S. Barocas, A. D. Selbst, and M. Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), New York, NY, USA, 2020. ACM.

[4] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317–331, 2018.

[5] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. Compressed sensing using generative models. In ICML. PMLR, 2017.

[6] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069, 2018.

[7] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373, 2017.

[8] S. Dandl, C. Molnar, M. Binder, and B. Bischl. Multi-objective counterfactual explanations. In International Conference on Parallel Problem Solving from Nature, pages 448–469. Springer, 2020.

[9] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Advances in Neural Information Processing Systems (NeurIPS), 2018.

[10] R. Dominguez-Olmedo, A. H. Karimi, and B. Schölkopf. On the adversarial robustness of causal algorithmic recourse. In International Conference on Machine Learning, pages 5324–5342. PMLR, 2022.

[11] R. Dominguez-Olmedo, A.-H. Karimi, and B. Schölkopf. On the adversarial robustness of causal algorithmic recourse. arXiv:2112.11313, 2021.

[12] E. Erdemir, J. Bickford, L. Melis, and S. Aydore. Adversarial robustness with non-uniform perturbations. Advances in Neural Information Processing Systems, 34:19147–19159, 2021.

[13] D. Garreau and U. Luxburg. Explaining the explainer: A first theoretical analysis of lime. In International Conference on Artificial Intelligence and Statistics (AISTATS), pages 1287–1296. PMLR, 2020.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[15] R. Hamon, H. Junklewitz, and I. Sanchez. Robustness and explainability of artificial intelligence. Publications Office of the European Union, 2020.

[16] M. Hardt and T. Ma. Identity matters in deep learning. In International Conference on Learning Representations (ICLR), 2017.

[17] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pages 43–58, 2011.

[18] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. Model-agnostic counterfactual explanations for consequential decisions. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

[19] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050, 2020.

[20] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Computing Surveys (CSUR), 2021.

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

[22] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In Conference on Neural Information Processing Systems (NeurIPS), 2020.

[23] Z. Kolter and A. Madry. Adversarial robustness - theory and practice. https://adversarial-ml-tutorial.org/linear_models/ .

[24] A. Kurakin, I. Goodfellow, S. Bengio, et al. Adversarial examples in the physical world, 2016.

[25] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. Inverse classification for comparison-based interpretability in machine learning. arXiv, 2017.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[27] D. Mahajan, C. Tan, and A. Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277, 2019.

[28] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, and H. Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2022.

[29] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, and H. Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In AISTATS. PMLR, 2022.

[30] M. Pawelczyk, K. Broelemann, and G. Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In Proceedings of The Web Conference 2020 (WWW). ACM, 2020.

[31] M. Pawelczyk, K. Broelemann, and G. Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In Proceedings of The Web Conference 2020, pages 3126–3132, 2020.

[32] M. Pawelczyk, T. Datta, J. van-den Heuvel, G. Kasneci, and H. Lakkaraju. Algorithmic recourse in the face of noisy human responses. arXiv preprint arXiv:2203.06768, 2022.

[33] K. Rawal, E. Kamar, and H. Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. arXiv:2012.11788, 2021.

[34] K. Rawal and H. Lakkaraju. Interpretable and interactive summaries of actionable recourses. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020.

[35] E. Rosenfeld, E. Winston, P. Ravikumar, and Z. Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In International Conference on Machine Learning, pages 8230–8241. PMLR, 2020.

[36] H. Shah, P. Jain, and P. Netrapalli. Do input gradients highlight discriminative features? Advances in Neural Information Processing Systems, 34:2046–2059, 2021.

[37] A. Sidford. sidford_mse213_cs269o_lec4_v5_share. https://web.stanford.edu/~sidford/courses/20fa_opt_theory/sidford_2020fa_mse213_cs269o_lec4.pdf . (Accessed on 09/06/2022).

[38] D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. Counterfactual explanations can be manipulated. In Advances in Neural Information Processing Systems (NeurIPS), volume 34, 2021.

[39] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv, 2013.

[40] S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. NeurIPS, 2021.

[41] S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. In Advances in Neural Information Processing Systems (NeurIPS), volume 34, 2021.

[42] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), 2019.

[43] B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. Proceedings of the conference on fairness, accountability, and transparency, pages 10–19, 2019.

[44] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.

[45] A. Van Looveren and J. Klaise. Interpretable counterfactual explanations guided by prototypes. arXiv preprint arXiv:1907.02584, 2019.

[46] S. Venkatasubramanian and M. Alfano. The philosophical basis of algorithmic recourse. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), New York, NY, USA, 2020. ACM.

[47] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. arXiv:2010.10596, 2020.

[48] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing, 10(3152676):10–5555, 2017.

[49] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard JL & Tech., 2017.

[50] Z. Zhao, D. Dua, and S. Singh. Generating natural adversarial examples. arXiv preprint arXiv:1710.11342, 2017.

A Related Works

Algorithmic Recourse. Several approaches have been proposed in recent literature to provide recourses to affected individuals [9, 49, 42, 45, 30, 27, 18, 22, 8]. These approaches can be broadly categorized along the following dimensions [47]: type of the underlying predictive model (e.g., tree based vs. differentiable classifier), type of access they require to the underlying predictive model (e.g., black box vs. gradient access), whether they encourage sparsity in counterfactuals (i.e., only a small number of features should be changed), whether counterfactuals should lie on the data manifold, whether the underlying causal relationships should be accounted for when generating counterfactuals, and whether the output produced by the method should be multiple diverse counterfactuals or a single counterfactual. In addition, [34] also studied how to generate global, interpretable summaries of counterfactual explanations. Some recent works also demonstrated that the recourses output by state-of-the-art techniques might not be robust, i.e., small perturbations to the original input [11, 38] to the underlying model [41, 33], or the recourse [32] itself may render the previously prescribed recourse(s) invalid. These works also formulated and solved minimax optimization problems to construct robust recourses to address the aforementioned challenges.

Adversarial Examples and Robustness. Prior works have shown that complex machine learning models, such as deep neural networks, are vulnerable to small changes in input [39]. This behavior of predictive models allows for generating adversarial examples (AEs) by adding in nitesimal changes to input targeted to achieve adversary-selected outcomes [39, 14]. Prior works have proposed several techniques to generate AEs using varying degrees of access to the model, training data, and the training procedure [6]. While gradient-based methods [14, 24] return the smallest input perturbations which flip the label as adversarial examples, generative methods constrain the search for adversarial examples to the training data-manifold. Finally, some methods generate

adversarial examples for non-differentiable and non-decomposable measures in complex domains such as speech recognition and image segmentation. Prior works have shown that Empirical Risk Minimization (ERM) does not yield models that are robust to adversarial examples [14, 24]. Hence, to reliably train adversarially robust models, [25] proposed the adversarial training objective which minimizes the worst-case loss within some small perturbation region around the input instances.

Intersections between Adversarial ML and Model Explanations. There has been a growing interest in studying the intersection of adversarial ML and model explainability [15, 16, 29]. Among all the existing works focusing on this intersection, two explorations are relevant to our work. [36] studied the interplay between adversarial robustness and post hoc explanations, and [36] demonstrated that gradient-based feature attribution methods (e.g., vanilla gradients, gradient times input, integrated gradients, smoothgrad) may severely violate the primary assumption of attribution – features with higher attribution are more important for model prediction – in case of non-robust models. However, their results also demonstrate that such a violation does not occur when the underlying models are robust to ℓ_2 and ℓ_{inf} input perturbations. More recently, [28] demonstrated that recourses generated by certain state-of-the-art methods are very similar to adversarial examples, and also argued that the methods proposed to output recourses and adversarial examples are designed with similar goals of changing the input minimally in order to achieve the desired outcome. While the aforementioned works explored the connections between adversarial ML and model explanations, none of these works focus on analyzing the impact of adversarially robust models on the recourses output by state-of-the-art algorithms.

B Proof for Theorems in Section 3

Here, we provide detailed proofs of the Lemmas and Theorems defined in Section 3.

B.1 Proof for Theorem 1

Theorem 1. (Cost difference for SCFE) For a given instance x, let x^0_{NR} and x^0_R be the recourse generated using Wachter's algorithm for the non-robust and adversarially robust models. Then, for a normalized Lipschitz activation function $\phi(\cdot)$, the difference in the recourse for both models can be bounded as:

$$\frac{}{+1}\,'(s; x; w_{NR}, w_R) \quad k \quad {}_{NR} \quad {}_R k_2 \quad \frac{}{+ k w_{NR} k^2}\, k w_{NR} k + \frac{}{+ k w_R k^2}\, k w_R k; \qquad (10)$$

where w_{NR} and w_R are the weights of the non-robust and adversarially robust models, is the regularization coefficient in Wachter's algorithm, $'(\cdot)$ is a function that measures the shift in the model weights using the target and predicted scores.

Proof. Following the definition of SCFE in Equation 2, we can find a counterfactual sample x^0 that is "closest" to the original instance x by minimizing the following objective:

$$\arg\min_{x^0} (h(x^0) \quad y^0)^2 + \quad d(x^0, x); \qquad (11)$$

where s is the target score, is the regularization coefficient, and $d(\cdot)$ is the distance between the original and counterfactual sample x^0.

Lower bound. Using Lemma 1, the optimal cost for generating a valid recourse for a non-robust ($_{NR}$) and adversarially robust ($_R$) model can be written as:

$$_{NR} = m_{NR} \frac{}{+ k w_{NR} k_2^2} \quad w_{NR} \qquad (12)$$

$$_R = m_R \frac{}{+ k w_R k_2^2} \quad w_R; \qquad (13)$$

where $m_{NR} = s \quad w_{NR}^T x$ and $m_R = s \quad w_R^T x$.

Subtracting and taking ℓ_2-norm on both sides of Eqn. 12 and Eqn. 13, we get:

$$\|w_{NR} - w_R\|_2 = \left\| \frac{m_{NR}}{+ \|w_{NR}\|_2^2} w_{NR} - \frac{m_R}{+ \|w_R\|_2^2} w_R \right\|_2$$

$$\|w_{NR} - w_R\|_2 \geq \left| \left\| \frac{m_{NR}}{+ \|w_{NR}\|_2^2} w_{NR} \right\|_2 - \left\| \frac{m_R}{+ \|w_R\|_2^2} w_R \right\|_2 \right|$$

(Using reverse triangle inequality)

$$\|w_{NR} - w_R\|_2 \geq \left| m_{NR} \frac{1}{+ \|w_{NR}\|_2^2} \|w_{NR}\|_2 - m_R \frac{1}{+ \|w_R\|_2^2} \|w_R\|_2 \right|$$

$$\|w_{NR} - w_R\|_2 \geq \left| m_{NR} \frac{1}{\|w_{NR}\|_2^2} \|w_{NR}\|_2 - m_R \frac{1}{\|w_R\|_2^2} \|w_R\|_2 \right| \quad \text{(Since, } \ll \|w\|_2\text{)}$$

$$\|w_{NR} - w_R\|_2 \geq \left| m_{NR} \frac{1}{\|w_{NR}\|_2^2} \|w_{NR}\|_2 - m_R \frac{1}{\|w_R\|_2^2} \|w_R\|_2 \right|$$

$$\|w_{NR} - w_R\|_2 \geq \left| \frac{m_{NR}}{\|w_{NR}\|_2} - \frac{m_R}{\|w_R\|_2} \right|$$

Upper bound. Again, using the optimal recourse cost (Definition 1), we can derive the upper bound of the cost difference for generating recourses using non-robust and adversarially robust models:

$$\begin{aligned} \|w_{NR} - w_R\|_2 &= \frac{(s - w_{NR}^T x)}{+ \|w_{NR}\|^2} w_{NR} - \frac{(s - w_R^T x)}{+ \|w_R\|^2} w_R \\ &= \frac{(s - w_{NR}^T x)}{+ \|w_{NR}\|^2} w_{NR} + \frac{(w_R^T x - s)}{+ \|w_R\|^2} w_R \\ &\leq \left| \frac{(s - w_{NR}^T x)}{+ \|w_{NR}\|^2} \right| \|w_{NR}\| + \left| \frac{(s - w_R^T x)}{+ \|w_R\|^2} \right| \|w_R\| \quad \text{(Using Triangle Inequality)} \end{aligned}$$

Note that the difference between the target and the predicted score for both non-robust and adversarially robust models is upper bounded by a term that is always positive. Hence, we get:

$$\|w_{NR} - w_R\|_2 \leq \frac{}{+ \|w_{NR}\|^2} \|w_{NR}\| + \frac{}{+ \|w_R\|^2} \|w_R\|$$

□

B.2 Proof for Theorem 2

Theorem 2. (Cost difference for C-CHVAE) Let z_{NR} and z_R be the solution returned by the C-CHVAE [31] algorithmic recourse method by sampling from ℓ_p-norm ball in the latent space using an L-Lipschitz generative model $G(\cdot)$ for a non-robust and adversarially robust model. By definition of the recourse method, let $x_{NR} = G(z_{NR})$ and $x_R = G(z_R)$ be the corresponding recourses in the input space. The difference between them can then be bounded as:

$$L(r_R - r_{NR}) \leq \|x_R - x_{NR}\|_p \leq L(r_R + r_{NR}); \tag{14}$$

where L is the Lipschitz constant of the generative model, and r_{NR} and r_R be the corresponding radii chosen by the algorithm such that they successfully return a recourse for the non-robust and adversarially robust model.

Proof. From the formulation of the counterfactual algorithm, we can write the difference between x_R and x_{NR} as:

$$\|x_R - x_{NR}\|_p = \|G(z_R) - G(z_{NR})\|_p \tag{15}$$

Lower bound. Here, we present a lower bound on the ℓ_p-norm of the cost difference between a baseline and robust model. Using Equation 15, we get:

$$\|x_R - x_{NR}\|_p = \|G(z_R) - G(z) + G(z) - G(z_{NR})\|_p \tag{16}$$

$$\|G(z_R) - G(z)\|_p - \|G(z_{NR}) - G(z)\|_p \quad \text{(since } \|a-b\|_p \geq \|a\|_p - \|b\|_p\text{)}$$

$$\geq L\|z_R - z\|_p - L\|z_{NR} - z\|_p \tag{17}$$

$$\|x_R - x_{NR}\|_p \geq L(r_R - r_{NR}); \quad \text{(Using [37])}$$

where z is the latent space representation for the original point x, and r_R and r_{NR} are the radius of the ℓ_p-norm for generating samples from the robust and baseline model. Note that using the radius of the ℓ_p norm in the above equation provides a tighter lower bound.

Upper bound. Using Equation 15, we can derive the upper bound using Lemma 1 and the triangle inequality.

$$\|x_R - x_{NR}\|_p \leq \|G(z_R) - x\|_p + \|x - G(z_{NR})\|_p \quad \text{(Using triangle inequality)}$$

$$= \|G(z_R) - G(z)\|_p + \|G(z) - G(z_{NR})\|_p \tag{18}$$

$$\leq L\|z_R - z\|_p + L\|z - z_{NR}\|_p \quad \text{(Using Lemma 1)}$$

$$\|x_R - x_{NR}\|_p \leq L(r_R + r_{NR}); \tag{19}$$

where r_R and r_{NR} is the radius of the ℓ_p-norm for generating samples from the robust and baseline model, respectively. □

B.3 Proof for Lemma 1

Lemma 1. (Difference between non-robust and adversarially robust model weights) For a given instance x, let w_{NR} and w_R be weights of the non-robust and adversarially robust model. Then, for a normalized Lipschitz activation function $\phi(\cdot)$, the difference in the weights w can be bounded as:

$$\|\Delta w\|_2 \leq n\eta(y\|x^T\|_2 + \sqrt{p}\epsilon) \tag{20}$$

where η is the learning rate, ϵ is the ℓ_2-norm perturbation ball, y is the label for x, n is the total number of training epochs, and d is the dimension of the input features.

Proof. Without loss of generality, we consider the case of binary classification which uses the binary cross entropy or logistic loss. Let us denote the baseline and robust models as $f_{NR}(x) = w_{NR}^T x$ and $f_R(x) = w_R^T x$, where we have removed the bias term for simplicity. We consider the class label as $y \in \{+1, -1\}$, and loss function $L(f(x)) = \log(1 + \exp(-y \cdot f(x)))$. Note that an adversarially robust model $f_R(x)$ is commonly trained using a min-max objective, where the inner maximization problem is given by:

$$\max_{\|\delta\| \leq \epsilon} L(w_R^T(x + \delta); y); \tag{21}$$

where δ is the adversarial perturbation added to a given sample x and ϵ denotes the the perturbation norm ball around x. Since our loss function is monotonic decreasing, the maximization of the loss function applied to a scalar is equivalent to just minimizing the scalar quantity itself, i.e.,

$$\max_{\|\delta\| \leq \epsilon} L(y(w_R^T(x + \delta))) = L(\min_{\|\delta\| \leq \epsilon} y(w_R^T(x + \delta))) = L(y(w_R^T x) + \min_{\|\delta\| \leq \epsilon} y w_R^T \delta) \tag{22}$$

The optimal solution to $\min_{\|\delta\| \leq \epsilon} y w_R^T \delta$ is given by $-\epsilon\|w_R^T\|_1$ [23]. Therefore, instead of solving the min-max problem for an adversarially robust model, we can convert it to a pure minimization problem, i.e.,

$$\min_{w_R} L(y(w_R^T x) - \epsilon\|w_R\|_1 \tag{23}$$

Correspondingly, the minimization objective for a baseline model is given by $\min_{w_{NR}} L(y(w_{NR}^T x))$. Looking into the training dynamics under gradient descent, we can define the weights at epoch `t' for a baseline and robust model as a function of the Jacobian of the loss function with respect to their corresponding weights, i.e.,

$$\frac{w_{NR} - w_0}{} = \eta \nabla_{w_{NR}} L(-y \cdot f_{NR}(x)); \tag{24}$$

$$\underline{w_R \quad w_0} = \; r_{w_R} L_R(y{:}f_R(x) \quad kw_R k_1); \tag{25}$$

where $\;$ is the learning rate of the gradient descent optimizer, w_0 is the weight initialization of both models.

$$r_{w_{NR}} L_{NR}(y{:}f_{NR}(x)) = \frac{\exp(\;\; y{:}f_{NR}(x))}{1+\exp(\;\; y{:}f_{NR}(x))} y{:}x^T$$

$$r_{w_R} L_R(y{:}f_R(x) \quad kw_R k_1) = \frac{\exp(\;\; y{:}f_R(x) + \;\; kw_R k_1)}{1+\exp(\;\; y{:}f_R(x) + \;\; kw_R k_1)}(\;\; y{:}x^T + \;\; {:}\mathrm{sign}(w_R));$$

where $\mathrm{sign}(x)$ return +1, -1, 0 for $x > 0$, $x < 0$, $x = 0$ respectively and $(x) = \frac{1}{1+\exp(\;\; x)}$ is the sigmoid function. Let us denote the weights of the baseline and robust models at the n-th iteration as w_{NR}^n and w_R^n, respectively. Hence, we can define the n-th step of the gradient-descent for both models as:

$$w_{NR}^n \quad w_{NR}^{n-1} = r_{w_{NR}^{n-1}} L_{NR}(\;) \tag{26}$$

$$w_R^n \quad w_R^{n-1} = r_{w_R^{n-1}} L_R(\;); \tag{27}$$

where $\;$ is the learning rate of the gradient descent optimizer. Taking $n = 1$, we get:

$$w_{NR}^1 \quad w_{NR}^0 = r_{w_{NR}^0} L_{NR}(\;) \tag{28}$$

$$w_R^1 \quad w_R^0 = r_{w_R^0} L_R(\;); \tag{29}$$

where w_{NR}^0 and w_R^0 are the same initial weights for the baseline and robust models. Subtracting both equations, we get:

$$\underline{w_{NR}^1 \quad w_R^1} = r_{w_{NR}^0} L_{NR}(\;) \quad r_{\quad w_R^0} L_R(\;) \tag{30}$$

Similarly, for $n = 2$ and using Equation 30, we get the following relation:

$$\underline{w_{NR}^2 \quad w_R^2} \quad \underline{w_{NR}^1 \quad w_R^1} = r_{w_{NR}^1} L_{NR}(\;) \quad r_{\quad w_R^1} L_R(\;)$$

$$\underline{w_{NR}^2 \quad w_R^2} = r_{w_{NR}^0} L_{NR}(\;) + r_{w_{NR}^1} L_{NR}(\;) \quad r_{\quad w_R^0} L_R(\;) \quad r_{\quad w_R^1} L_R(\;)$$

Using the above equations, we can now write the difference between the weights of the baseline and robust models at the n-th iteration as:

$$\begin{aligned}
\underline{w_{NR}^n \quad w_R^n} &= \sum_{i=0}^{n-1} r_{w_{NR}^i} L_{NR}(\;) \quad \sum_{i=0}^{n-1} r_{w_R^i} L_R(\;) \\
&= \sum_{i=0}^{n-1} (\;(y{:}f_{NR}^i(x)) \quad 1) y{:}x^T \quad \sum_{i=0}^{n-1} (\;(y{:}f_R^i(x) \quad \|w_R^i\|_1) \quad 1)(y{:}x^T \quad \mathrm{sign}(w_R^i)) \\
&= \sum_{i=0}^{n-1} (y{:}f_{NR}^i(x)) y{:}x^T \quad \sum_{i=0}^{n-1} (y{:}f_R^i(x) \quad \|w_R^i\|_1)(y{:}x^T \quad \mathrm{sign}(w_R^i)) + \mathrm{sign}(w_R^i) \\
&\quad \sum_{i=0}^{n-1} (y{:}f_{NR}^i(x)) y{:}x^T \quad \sum_{i=0}^{n-1} (y{:}f_R^i(x))(y{:}x^T \quad \mathrm{sign}(w_R^i)) + \mathrm{sign}(w_R^i) \\
&\qquad\qquad \text{(Using } (a \quad b) \quad (a) \text{ for } b > 0) \\
&\quad \sum_{i=0}^{n-1} (y{:}f_{NR}^i(x) \quad (y{:}f_R^i(x)) \; y{:}x^T + \sum_{i=0}^{n-1} (y{:}f_R^i(x)) \quad 1 \quad \mathrm{sign}(w_R^i)
\end{aligned}$$

Using ℓ_2-norm on both sides, we get:

$$\begin{aligned}
&\frac{1}{-}k\mathbf{w}_{\mathrm{NR}}^{n} \quad \mathbf{w}_{\mathrm{R}}^{n}k_2 \quad k\sum_{i=0}^{n-1} (y{:}f_{\mathrm{NR}}^{i}(\mathbf{x}) \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \; y{:}\mathbf{x}^T + \sum_{i=0}^{n-1} (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1 \quad \mathrm{sign}(\mathbf{w}_R^i)k_2 \\
&\quad k\sum_{i=0}^{n-1} (y{:}f_{\mathrm{NR}}^{i}(\mathbf{x}) \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \; y{:}\mathbf{x}^T k_2 + k\sum_{i=0}^{n-1} (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1 \quad \mathrm{sign}(\mathbf{w}_R^i)k_2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using Triangle Inequality)} \\
&\quad k\sum_{i=0}^{n-1} (y{:}f_{\mathrm{NR}}^{i}(\mathbf{x}) \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \; y{:}\mathbf{x}^T k_2 + \bar{p}_d \, k\sum_{i=0}^{n-1} k \; (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1k_2 \\
&\quad \sum_{i=0}^{n-1} k \quad (y{:}f_{\mathrm{NR}}^{i}(\mathbf{x}) \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \; y{:}\mathbf{x}^T k_2 + \bar{p}_d \, k\sum_{i=0}^{n-1} k \; (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1k_2 \\
&\quad \sum_{i=0}^{n-1} k \quad (y{:}f_{\mathrm{NR}}^{i}(\mathbf{x})) \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \; k_2 k y{:}\mathbf{x}^T k_2 + \bar{p}_d \, k\sum_{i=0}^{n-1} k \; (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1k_2 \\
&\quad nky{:}\mathbf{x}^T k_2 + \bar{p}_d \, k\sum_{i=0}^{n-1} k \; (y{:}f_{\mathrm{R}}^{i}(\mathbf{x})) \quad 1k_2 \\
&\quad nky{:}\mathbf{x}^T k_2 + \bar{p}_d \, k\sum_{i=0}^{n-1} (1 \quad (y{:}f_{\mathrm{R}}^{i}(\mathbf{x}))) \quad \text{(since the term inside } k\,k_2 \text{ is a scalar)} \\
&\quad nky{:}\mathbf{x}^T k_2 + \bar{p}_d n \\
k \quad \mathbf{w}k_2 \quad n \; (yk\mathbf{x}^T k_2 + \bar{p}_d)
\end{aligned}$$

□

B.4 Validity

Theorem 3. *(Validity Comparison) For a given instance* $\mathbf{x} \in \mathbb{R}^d$ *and desired target label denoted by unity, let* $\mathbf{x}_{\mathrm{R}}$ *and* $\mathbf{x}_{\mathrm{NR}}$ *be the counterfactuals for adversarially robust* $f_{\mathrm{R}}(\mathbf{x})$ *and non-robust* $f_{\mathrm{NR}}(\mathbf{x})$ *models respectively. Then,* $Pr(f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{NR}}) = 1) \quad Pr(f_{\mathrm{R}}(\mathbf{x}_{\mathrm{R}}) = 1)$ *if* $|f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{R}}) \quad f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{NR}})| \; n \; (yk\mathbf{x}^T k_2 + \bar{p}_d)k\mathbf{x}_{\mathrm{R}}k$*, where* $\quad$ *is the learning rate,* $\quad$ *is the* ℓ_2*-norm perturbation ball,* y *is the label for* $\mathbf{x}$*, and* n *is the total number of training epochs.*

Proof. In a logistic regression case, $Pr(f(\mathbf{x}) = 1) = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1+e^{\mathbf{w}^T\mathbf{x}}}$, which is the sigmoid of the model output. Next, we derive the difference in probability of a valid recourse from non-robust and adversarially robust model:

$$Pr(f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{NR}}) = 1) \quad Pr(f_{\mathrm{R}}(\mathbf{x}_{\mathrm{R}}) = 1) = \frac{e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}}}{1 + e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}}} \quad \frac{e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}}}{1 + e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}}} \tag{31}$$

$$= \frac{e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}} \quad e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}}}{(1 + e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}})(1 + e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}})} \tag{32}$$

Since $(1 + e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}})(1 + e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}}) > 0$, so $Pr(f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{NR}}) = 1) \quad Pr(f_{\mathrm{R}}(\mathbf{x}_{\mathrm{R}}) = 1))$ occurs when,

$$e^{\mathbf{w}_{\mathrm{NR}}^T\mathbf{x}_{\mathrm{NR}}} \quad e^{\mathbf{w}_{\mathrm{R}}^T\mathbf{x}_{\mathrm{R}}} \tag{33}$$

$$\mathbf{w}_{\mathrm{NR}}^T(\mathbf{x}_{\mathrm{NR}} \quad \mathbf{x}_{\mathrm{R}}) \quad (\mathbf{w}_{\mathrm{R}}^T \quad \mathbf{w}_{\mathrm{NR}}^T)\mathbf{x}_{\mathrm{R}} \qquad \text{(Taking natural logarithm on both sides)}$$

$$\mathbf{w}_{\mathrm{NR}}^T(\mathbf{x}_{\mathrm{R}} \quad \mathbf{x}_{\mathrm{NR}}) \quad (\mathbf{w}_{\mathrm{NR}}^T \quad \mathbf{w}_{\mathrm{R}}^T)\mathbf{x}_{\mathrm{R}} \tag{34}$$

$$\mathbf{w}_{\mathrm{NR}}^T(\mathbf{x}_{\mathrm{R}} \quad \mathbf{x}_{\mathrm{NR}}) \quad (\mathbf{w}_{\mathrm{NR}}^T \quad \mathbf{w}_{\mathrm{R}}^T)\mathbf{x}_{\mathrm{R}} \qquad \text{(Taking norm on both sides)}$$

$$\mathbf{w}_{\mathrm{NR}}^T(\mathbf{x}_{\mathrm{R}} \quad \mathbf{x}_{\mathrm{NR}}) \quad k\mathbf{w}_{\mathrm{NR}} \quad \mathbf{w}_{\mathrm{R}}kk\mathbf{x}_{\mathrm{R}}k \qquad \text{(Using Cauchy-Schwartz)}$$

$$\mathbf{w}_{\mathrm{NR}}^T(\mathbf{x}_{\mathrm{R}} \quad \mathbf{x}_{\mathrm{NR}}) \quad n \; (yk\mathbf{x}^T k_2 + \bar{p}_d)k\mathbf{x}_{\mathrm{R}}k \qquad \text{(From Lemma 1)}$$

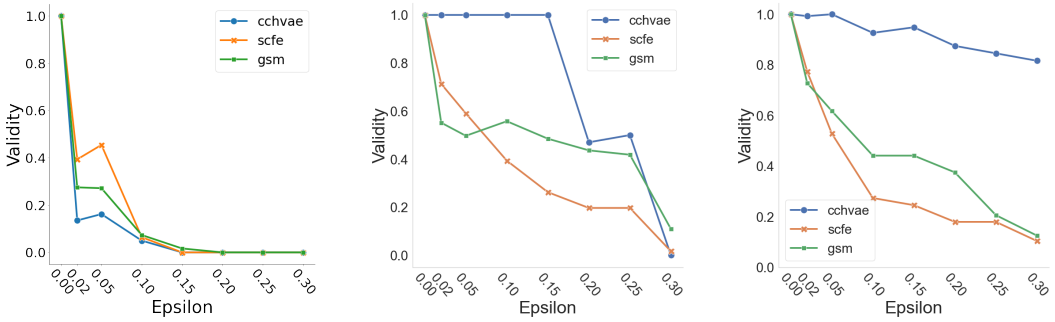
$$|f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{R}}) \quad f_{\mathrm{NR}}(\mathbf{x}_{\mathrm{NR}})| \quad n \; (yk\mathbf{x}^T k_2 + \bar{p}_d)k\mathbf{x}_{\mathrm{R}}k \tag{35}$$

□

C Additional Experimental Results

In this section, we have plots for cost differences, validity, and adversarial accuracy for the two logistic regression and neural network models trained on three real-world datasets.

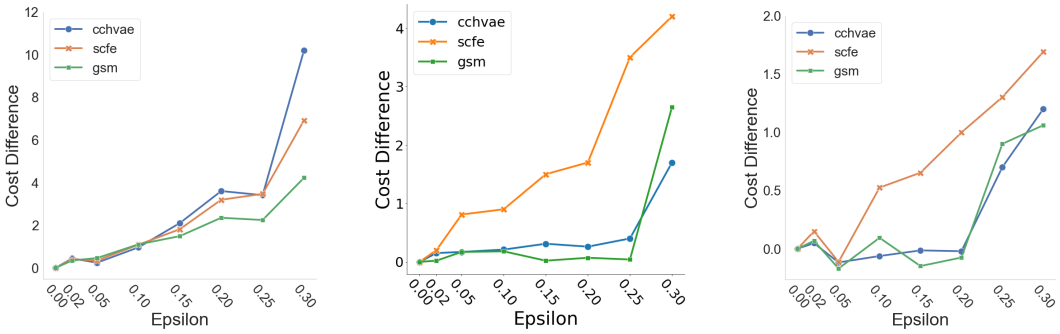
(a) NN – Adult (b) NN – German Credit (c) NN – COMPAS



(d) LR – Adult (e) LR – German Credit (f) LR – COMPAS

Figure 5: Analyzing validity of recourse generated using non-robust and adversarially robust Logistic Regression(LR) and Neural Networks (NN) for Adult, COMPAS, and German Credit datasets. We find that the validity decreases for increasing values of .

(a) NN – Adult (b) NN – German Credit (c) NN – COMPAS



(d) LR – Adult (e) LR – German Credit (f) LR – COMPAS

Figure 6: Analyzing cost differences between recourse generated using non-robust and adversarially robust Logistic Regression (LR) and Neural Networks(NN) for Adult, COMPAS, and German Credit datasets. We find that the cost difference (i.e., ℓ_2 norm) between the recourses generated for non-robust and adversarially robust models increases for increasing values of ϵ.