On the Trade-offs between Adversarial Robustness and Actionable Explanations

Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju

Harvard University
skrishna@g.harvard.edu

Abstract

As machine learning models are increasingly being employed in various high-stakes settings, it becomes important to ensure that predictions of these models are not only adversarially robust, but also readily explainable to relevant stakeholders. However, it is unclear if these two notions can be simultaneously achieved or if there exist trade-offs between them. In this work, we make one of the first attempts at studying the impact of adversarially robust models on actionable explanations which provide end users with a means for recourse. We theoretically and empirically analyze the cost (ease of implementation) and validity (probability of obtaining a positive model prediction) of recourses output by state-of-the-art algorithms when the underlying models are adversarially robust vs. non-robust. More specifically, we derive theoretical bounds on the differences between the cost and the validity of the recourses generated by state-of-the-art algorithms for adversarially robust vs. non-robust linear and non-linear models. Our empirical results with multiple real-world datasets validate our theoretical results and show the impact of varying degrees of model robustness on the cost and validity of the resulting recourses. Our analyses demonstrate that adversarially robust models significantly increase the cost and reduce the validity of the resulting recourses, thus shedding light on the inherent trade-offs between adversarial robustness and actionable explanations.

1 Introduction

In recent years, machine learning (ML) models have made significant strides, becoming indispensable tools in high-stakes domains such as banking, healthcare, and criminal justice. As these models continue to gain prominence, it is more crucial than ever to address the dual challenge of providing actionable explanations to individuals negatively impacted (e.g., denied loan applications) by model predictions, and ensuring adversarial robustness to maintain model integrity. Both prior research and recent regulations have emphasized the importance of adversarial robustness and actionable explanations, deeming them as key pillars of trustworthy machine learning (GDPR 2016; Hamon, Junklewitz, and Sanchez 2020; Voigt and Von dem Bussche 2017) that are critical to real-world applications.

Existing machine learning research has explored adversarial robustness and actionable explanations in different silos, where several techniques have been proposed for generating actionable explanations in practice (using counterfactual explanations) (Wachter, Mittelstadt, and Russell 2018; Ustun, Spangher, and Liu 2019; Pawelczyk, Broelemann, and Kasneci 2020; Karimi et al. 2020a) and adversarial examples (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014). The goal of a counterfactual explanation is to generate recourses for individuals impacted by model outcomes, e.g. when an individual is denied a loan by a predictive model, a counterfactual explanation informs them about which feature should be changed and by how much in order to obtain a positive outcome from the model. However, an adversarial attack aims to demonstrate the vulnerabilities of modern deep neural network (DNN) models and generates infinitesimal input perturbations to achieve adversary-selected model outcomes. Recently, adversarial training has been proposed as a defense against adversarial examples, with the goal of training robust models in adversarial scenarios (Madry et al. 2017). However, its potential impact on generating algorithmic recourse has not been studied yet, and this is the focus of our work.

Despite recent regulations (GDPR 2016; Hamon, Junklewitz, and Sanchez 2020) emphasizing the importance of adversarial robust models and generating actionable explanations for end-users, there has been little to no work in understanding the connection between these contrasting trustworthy properties. While previous works (Xie et al. 2020) study the trade-offs between adversarial robustness and accuracy on the original data, the trade-offs between adversarial robustness and recourse have been unexplored, and ours is one of the first works to shed light on these trade-offs across diverse models and datasets. To this end, only a few works (Bansal, Agarwal, and Nguyen 2020; Shafahi et al. 2019) show the impact of robust models on model gradients, where they find that input gradients are less noisier for adversarially robust models as compared to non-robust models. While these findings highlight the impact of robust models on model gradients, we aim to show that adversarial robustness does not come for free, and there are trade-offs between robustness and recourse costs.

Present work. In this study, we address the aforementioned gaps by presenting the first-ever investigation of the impact of adversarially robust models on algorithmic recourse. We provide a theoretical and empirical analysis of the *cost* (ease of implementation) and *validity* (likelihood of achieving a desired model prediction) of the recourses generated by state-of-the-art algorithms for adversarially robust and non-robust models. In particular, we establish theoretical bounds on the differences in cost and validity for recourses produced by various gradient-based (Laugel et al. 2017; Wachter, Mittelstadt, and Russell 2018) and manifold-based (Pawelczyk, Broelemann, and Kasneci 2020) recourse methods for adversarially robust and non-robust linear and non-linear models (see Section 4). To achieve this, we first derive theoretical bounds on the differences between the weights (parameters) of adversarially robust and non-robust linear and non-linear models, and then use these bounds to establish the differences in cost and validity of the corresponding recourses. It is important to note that in this study, we analyze the impact on algorithmic recourses when the model is made adversarially robust. This should not be confused with the analysis comparing adversarial examples and algorithmic recourses(Freiesleben 2022; Pawelczyk, Broelemann, and Kasneci 2020).

We conducted extensive experiments with multiple real-world datasets from diverse domains (Section 5). Our theoretical and empirical analyses provide several interesting insights into the relationship between adversarial robustness and algorithmic recourse: i) the cost of recourse increases with the degree of robustness of the underlying model, and ii) the validity of recourse deteriorates as the degree of robustness of the underlying model increases. Additionally, we conducted a qualitative analysis of the recourses generated by state-of-the-art methods, and observed that the number of valid recourses for any given instance decreases as the underlying model's robustness increases. More broadly, our analyses and findings shed light on the the inherent trade-offs between adversarial robustness and actionable explanations.

2 Related Work

Algorithmic Recourse. Several approaches have been proposed in recent literature to provide recourses to affected individuals (Wachter, Mittelstadt, and Russell 2018; Ustun, Spangher, and Liu 2019; Van Looveren and Klaise 2019; Pawelczyk, Broelemann, and Kasneci 2020; Mahajan, Tan, and Sharma 2019; Karimi et al. 2020a,c). These approaches can be broadly categorized along the following dimensions (Verma, Dickerson, and Hines 2020): *type of the underlying predictive model* (e.g., tree-based vs. differentiable classifier), *type of access* of the underlying predictive model (e.g., black box vs. gradient access), whether they encourage *sparsity* in counterfactuals (i.e., allowing changes in a small number of features), whether counterfactuals should lie on the *data manifold*, whether the underlying *causal relationships* should be accounted for when generating counterfactuals, and whether the produced output by the method should be *multiple diverse counterfactuals* or a single counterfactual. Recent works also demonstrate that recourses output by state-of-the-art techniques are not robust, i.e., small perturbations to the original instance (Dominguez-Olmedo, Karimi, and Schölkopf 2021; Slack et al. 2021), the underlying model (Upadhyay, Joshi, and Lakkaraju 2021; Rawal, Kamar, and Lakkaraju 2021), or the recourse (Pawelczyk et al. 2022b) itself may render the previously prescribed recourse(s) invalid. These works also proposed minimax optimization problems to find *robust* recourses to address the aforementioned challenges.

Adversarial Examples and Robustness. Prior works have shown that complex machine learning model, such as deep neural networks, are vulnerable to small changes in input (Szegedy et al. 2013). This behavior of predictive models allows for generating adversarial examples (AEs) by adding infinitesimal changes to input targeted to achieve adversary-selected outcomes (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014). Prior works have proposed several techniques to generate AEs using varying degrees of access to the model, training data, and the training procedure (Chakraborty et al. 2018). While gradient-based methods (Goodfellow, Shlens, and Szegedy 2014; Kurakin et al. 2016) return the smallest input perturbations which flip the label as adversarial examples, generative methods (Zhao, Dua, and Singh 2017) constrain the search for adversarial examples to the training data-manifold. Finally, some methods (Cisse et al. 2017) 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 (Goodfellow, Shlens, and Szegedy 2014; Kurakin et al. 2016). Hence, to reliably train adversarially robust models, Madry et al. (2017) 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 (Hamon, Junklewitz, and Sanchez 2020). Among all these works, two are relevant to our work (Pawelczyk et al. 2022a; Shah, Jain, and Netrapalli 2021). Shah, Jain, and Netrapalli (2021) studied the interplay between adversarial robustness and post hoc explanations (Shah, Jain, and Netrapalli 2021), demonstrating that gradient-based explanations violate the primary assumption of attributions – features with higher attribution are more important for model prediction – in case of non-robust models. Further, they show that such a violation does not occur when the underlying models are robust to ℓ_2 and ℓ_∞ input perturbations. More recently, Pawelczyk et al. (2022a) demonstrated that the distance between the recourses generated by state-of-the-art methods and adversarial examples is small for linear models. While existing works explore the connections between adversarial ML and model explanations, none focus on the trade-offs between adversarial robustness and actionable explanations, which is the focus of our work.

3 Preliminaries

Notation In this work, we denote a model $f : \mathbb{R}^d \rightarrow \mathbb{R}$, where $\mathbf{x} \in \mathcal{X}$ is a d-dimensional input sample, $\mathcal{X}$ is the training dataset, and the model is parameterized with weights $\mathbf{w}$. In addition, we represent the non-robust and adversarially robust models using $f_{\text{NR}}(\mathbf{x})$ and $f_{\text{R}}(\mathbf{x})$, and the linear and neural network models using $f^{\mathrm{L}}(\mathbf{x})$ and $f^{\mathrm{NTK}}(\mathbf{x})$. Below, we provide a brief overview of adversarially robust models, and some popular methods for generating recourses.

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 infinitesimal perturbations to the original samples targeted to change prediction label (Agarwal, Nguyen, and Schonfeld 2019). One standard approach to ameliorate this problem is via adversarial training which minimizes the worst-case loss within some perturbation region (the perturbation model) (Kolter and Madry 2023). In particular, for a model f parameterized by weights $\mathbf{w}$, loss function $\ell(\cdot)$, and training data $\{\mathbf{x}_i, y_i\}_{i=\{1,2,\ldots,n\}} \in \mathcal{D}_{\text{train}}$, the optimization problem of minimizing the worst-case loss within ℓ_p-norm perturbation with radius ϵ is:

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

where $\mathcal{D}_{\text{train}}$ denotes the training dataset and $\Delta_{p,\epsilon} = \{\delta : \|\delta\|_p \leq \epsilon\}$ is the ℓ_p ball with radius ϵ centered around sample $\mathbf{x}$. We use $p = \infty$ for our theoretical analysis resulting in a closed-form solution of the model parameters $\mathbf{w}$.

Algorithmic Recourse. One way to generate recourses is by explaining to affected individuals what features in their profile need to change and by how much in order to obtain a positive outcome. Counterfactual explanations that essentially capture the aforementioned information can be therefore used to provide recourses. The terms *counterfactual explanations* and *algorithmic recourse* have, in fact, become synonymous in recent literature (Karimi et al. 2020b; Ustun, Spangher, and Liu 2019; Venkatasubramanian and Alfano 2020). In particular, methods that try to find algorithmic recourses do so by finding a counterfactual $\mathbf{x}' = \mathbf{x} + \zeta$ that is closest to the original instance $\mathbf{x}$ and change the model's prediction $f(\mathbf{x} + \zeta)$ to the target label, where ζ determines a set of changes that can be made to $\mathbf{x}$ in order to reverse the negative outcome. Next, we describe three popular recourse methods we analyze to understand the implications of adversarially robust models on algorithmic recourses.

Score CounterFactual Explanations (SCFE). For a given model $f : \mathbb{R}^d \to \mathbb{R}$, a distance function $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, and sample $\mathbf{x}$, Wachter, Mittelstadt, and Russell (2018) define the problem of generating a counterfactual $\mathbf{x}'{=}\mathbf{x} + \zeta$ using the following objective:

$$\arg\min_{\mathbf{x}'} \; (f(\mathbf{x}') - s)^2 + \lambda d(\mathbf{x}', \mathbf{x}), \qquad (2)$$

where s is the target score for the counterfactual $\mathbf{x}'$, λ is the regularization coefficient, and $d(\cdot)$ is the distance between sample $\mathbf{x}$ and its counterfactual $\mathbf{x}'$.

C-CHVAE. Given a Variational AutoEncoder (VAE) model with encoder $\mathcal{I}_\gamma$ and decoder $\mathcal{G}_\theta$ trained on the original data distribution $\mathcal{D}_{\text{train}}$, C-CHVAE (Pawelczyk, Broelemann, and Kasneci 2020) aims to generate recourses in the latent space $\mathcal{Z}$, where $\mathcal{I}_\gamma : \mathcal{X} \to \mathcal{Z}$. The encoder transforms a given sample $\mathbf{x}$ into a latent representation $\mathbf{z} \in \mathcal{Z}$ and the decoder takes $\mathbf{z}$ as input and generates $\hat{\mathbf{x}}$ as similar as possible to $\mathbf{x}$. Formally, C-CHVAE generates recourse using the following objective function:

$$\zeta^* = \arg\min_{\zeta\in\mathcal{Z}} \|\zeta\| \quad \text{such that} \quad f(\mathcal{G}_\theta(\mathcal{I}_\gamma(\mathbf{x}) + \zeta)) \neq f(\mathbf{x}), \qquad (3)$$

where ζ is the cost for generating a recourse, $\mathcal{I}_\gamma$ allows to search for counterfactuals in the data manifold and $\mathcal{G}_\theta$ projects the latent counterfactuals back to the input feature space.

Growing Spheres Method (GSM). While the above techniques directly optimize specific objective functions for generating counterfactuals, GSM (Laugel et al. 2017) uses a search-based algorithm to generate recourses by randomly sampling points around the original instance $\mathbf{x}$ until a sample with the target label is found. In particular, GSM method involves first drawing an ℓ_2-sphere around a given instance $\mathbf{x}$, randomly samples point within that sphere and checks whether any sampled points result in target prediction. This method then contracts or expands the sphere until a (sparse) counterfactual is found and returns it. GSM defines a minimization problem using a function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, where $c(\mathbf{x}, \mathbf{x}')$ is the cost of going from instance $\mathbf{x}$ to counterfactual $\mathbf{x}'$.

$$\mathbf{x}'^* = \arg\min_{\mathbf{x}'\in\mathcal{X}} \{c(\mathbf{x}, \mathbf{x}') \mid f(\mathbf{x}') \neq f(\mathbf{x})\}, \qquad (4)$$

where $\mathbf{x}'$ is sampled from the ℓ_2-ball around $\mathbf{x}$ such that $f(\mathbf{x}'){\neq}f(\mathbf{x})$, $c(\mathbf{x}, \mathbf{x}'){=}\|\mathbf{x}'-\mathbf{x}\|_2+\gamma\|\mathbf{x}'-\mathbf{x}\|_0$, and $\gamma \in \mathbb{R}_+$ is the weight associated to the sparsity objective.

4 Our Theoretical Analysis

Here, we perform a detailed theoretical analysis to bound the cost and validity differences of recourses generated by state-of-the-art methods when the underlying models are non-robust vs. adversarially robust, for linear and non-linear predictors. In particular, we compare the cost differences (Sec. 4.1) of the recourses obtained using 1) gradient-based methods like SCFE (Wachter, Mittelstadt, and Russell 2018) and 2) manifold-based methods like C-CHVAE (Pawelczyk, Broelemann, and Kasneci 2020). Finally, we show that the validity of the recourses generated using existing methods for robust models is lower compared to that of non-robust models (Sec. 4.2).

4.1 Cost Analysis

The cost of a generated algorithmic recourse is defined as the distance between the input instance $\mathbf{x}$ and the counterfactual $\mathbf{x}'$ obtained using a recourse finding method (Verma, Dickerson, and Hines 2020). Algorithmic recourses with lower costs are considered better as they achieve the desired outcome with minimal changes to input. Next, we theoretically analyze the cost difference of recourses generated for non-robust and adversarially robust linear and non-linear models. Below, we first find the weight difference between non-robust and adversarially robust models and then use these lemmas to derive the recourse cost differences.

Cost Analysis of recourses generated using SCFE method Here, we derive the lower and upper bound for the cost difference of recourses generated using SCFE (Wachter, Mittelstadt, and Russell 2018) method when the underlying models are non-robust vs. adversarially robust linear and non-linear models. We first derive a bound for the difference between non-robust and adversarially robust linear model weights.

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

$$\|\mathbf{w}_{\mathrm{NR}} - \mathbf{w}_{\mathrm{R}}\|_2 \leq \Delta \tag{5}$$

where $\Delta = N\eta(y\|\mathbf{x}^T\|_2 + \epsilon\sqrt{d})$*,* η *is the learning rate,* ϵ *is the* ℓ_2*-norm perturbation ball around the sample* $\mathbf{x}$*,* y *is the label for* $\mathbf{x}$*,* N *is the total number of training epochs, and* d *is the dimension of the input features. Subsequently, we show that* $\|\mathbf{w}_{\mathrm{NR}}\|_2 - \Delta \leq \|\mathbf{w}_{\mathrm{R}}\|_2 \leq \|\mathbf{w}_{\mathrm{NR}}\|_2 + \Delta$.

Proof Sketch. We separately derive the gradients for updating the weight for the non-robust and adversarially robust linear models. The proof uses sigmoidal and triangle inequality properties to derive the bound for the difference between the non-robust and adversarially robust linear model. In addition, we use reverse triangle inequality properties to show that the weights of the adversarially robust linear model are bounded by $\|\mathbf{w}_{\mathrm{NR}}\|_2 \pm \Delta$. See Appendix A.1 for detailed proof. [1] □

Implications: We note that the weight difference in Eqn. 1 is proportional to the ℓ_2-norm of the input and the square root of the number of dimensions of $\mathbf{x}$. In particular, the bound is tighter for samples with lower feature dimensions d and models with a smaller degree of robustness ϵ.

Next, we define the closed-form solution for the cost ζ^* to generate a recourse for the linear model.

Definition 1. *(Optimal cost for linear models (Pawelczyk et al. 2022a)) For a given scoring function* $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$*, the SCFE method generates a recourse for an input* $\mathbf{x}$ *using cost* ζ *such that:*

$$\zeta^* = m\frac{\lambda}{\lambda + \|\mathbf{w}\|_2^2} \cdot \mathbf{w}, \tag{6}$$

where $m = s - f(\mathbf{x}')$ *is the target residual,* s *is the target score for* $\mathbf{x}$*,* $\mathbf{w}$ *is the weight of the linear model, and* λ *is a given hyperparameter.*

We now derive the cost difference bounds of recourses generated using SCFE when the underlying model is non-robust and adversarially robust linear models.

Theorem 1. *(Cost difference of SCFE for linear models) For a given instance* $\mathbf{x}$*, let* $\mathbf{x}'_{\mathrm{NR}} = \mathbf{x} + \zeta_{\mathrm{NR}}$ *and* $\mathbf{x}'_{\mathrm{R}} = \mathbf{x} + \zeta_{\mathrm{R}}$ *be the recourse generated using Wachter's algorithm for the non-robust and adversarially robust linear models. Then, for a normalized Lipschitz activation function* $\sigma(\cdot)$*, the difference in the recourse for both models can be bounded as:*

$$\|\zeta_{\mathrm{NR}}\|_2 - \|\zeta_{\mathrm{R}}\|_2 \leq \left| \lambda \frac{2\|\mathbf{w}_{\mathrm{NR}}\|_2 + \Delta}{\|\mathbf{w}_{\mathrm{NR}}\|_2(\|\mathbf{w}_{\mathrm{NR}}\|_2 - \Delta)} \right|, \tag{7}$$

where $\mathbf{w}_{\mathrm{NR}}$ *is the weight of the non-robust model,* λ *is the scalar coefficient on the distance between original sample* $\mathbf{x}$ *and generated counterfactual* $\mathbf{x}'$*, and* Δ *is defined in Lemma 1.*

Proof Sketch. We use the optimal cost for recourses in linear models (see Def. 1) for deriving the cost difference bounds. The proof for the weight difference uses linear algebra and triangle inequality properties. See Appendix A.2 for the complete proof. □

Implications: The derived bounds imply that the differences between costs are a function of the quantity Δ (RHS term from Lemma 1), the weights of the non-robust model $||\mathbf{w}_{\mathrm{NR}}||_2$, and λ, where the bound of the difference between costs is tighter (lower) for smaller Δ values and when the ℓ_2-norm of the non-robust model weight is large (due to the quadratic term in the denominator). We note that the Δ term is a function of the ℓ_2-norm of the input $\mathbf{x}$ and the square root of the number of dimensions d of the input sample, where the bound is tighter for smaller feature dimensions d, models with a smaller degree of robustness ϵ, and $\mathbf{x}$ with larger ℓ_2-norms.

Next, we define the closed-form solution for the cost ζ^* required to generate a recourse when the underlying model is a wide neural network.

Definition 2. *(Kernel Matrix for ReLU networks (Du et al. 2019; Zhang and Zhang 2022)) The closed-form solution of the Neural Tangent Kernel for a two-layer neural network model with ReLU non-linear activation is given by:*

$$\mathbf{K}^{\infty}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\Big(\pi - \arccos(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\|\mathbf{x}_i\|\,\|\mathbf{x}_j\|})\Big)/2\pi, \tag{8}$$

where $\mathbf{K}^{\infty}$ *is the Neural Tangent Kernel matrix and* $\mathbf{x}_i \in \mathbb{R}^d$.

We now derive the difference between costs for generated SCFE recourses when the underlying model is non-robust vs. adversarially robust wide neural network model.

Theorem 2. *(Cost difference for SCFE for wide neural network) For an NTK model with weights* $\mathbf{w}_{\mathrm{NR}}^{\mathrm{NTK}}$ *and* $\mathbf{w}_{\mathrm{R}}^{\mathrm{NTK}}$ *for the non-robust and adversarially robust models, the cost difference between the recourses generated for sample* $\mathbf{x}$ *is bounded as:*

$$\|\zeta_{\mathrm{NR}}\|_2 - \|\zeta_{\mathrm{R}}\|_2 \leq \left| \frac{2}{\mathrm{H}(\|\bar{\mathbf{w}}_{\mathrm{NR}}^{\mathrm{NTK}}\|_2, \|\bar{\mathbf{w}}_{\mathrm{R}}^{\mathrm{NTK}}\|_2)} \right|, \tag{9}$$

where $\mathrm{H}(\cdot,\cdot)$ *denotes the harmonic mean,* $\bar{\mathbf{w}}_{\mathrm{NR}}^{\mathrm{NTK}} = \nabla_{\mathbf{x}}\mathbf{K}^{\infty}(\mathbf{x}, \mathbf{X})\mathbf{w}_{\mathrm{NR}}^{\mathrm{NTK}}$*,* $\mathbf{K}^{\infty}$ *is the NTK associated with the wide neural network model,* $\bar{\mathbf{w}}_{\mathrm{R}}^{\mathrm{NTK}} = \nabla_{\mathbf{x}}\mathbf{K}^{\infty}(\mathbf{x}, \mathbf{X}_{\mathrm{R}})\mathbf{w}_{\mathrm{R}}^{\mathrm{NTK}}$*,* $\mathbf{w}_{\mathrm{NR}}^{\mathrm{NTK}} = (\mathbf{K}^{\infty}(\mathbf{X}, \mathbf{X}) + \beta\mathbf{I}_n)^{-1}\mathbf{Y}$*,* $\mathbf{w}_{\mathrm{R}}^{\mathrm{NTK}} = (\mathbf{K}^{\infty}(\mathbf{X}_{\mathrm{R}}, \mathbf{X}_{\mathrm{R}}) + \beta\mathbf{I}_n)^{-1}\mathbf{Y}$*,* β *is the bias of the NTK model,* $(\mathbf{X}, \mathbf{X}_{\mathrm{R}})$ *are the training samples for the non-robust and adversarially robust models, and* $\mathbf{Y}$ *are the labels of the training samples.*

[1]Please refer to the arXiv link for appendix sections. arXiv link: https://arxiv.org/abs/2309.16452

Proof Sketch. The proof follows from Def. 2, where we use data processing, Taylor expansion, and triangle inequalities to bound the difference between costs of recourses output by SCFE for non-robust vs. adversarially robust wide neural network models. See Appendix A.3 for the complete proof. □

Implications: The proposed bounds imply that the difference in costs is bounded by the harmonic mean of the NTK models weights of non-robust and robust models, *i.e.,* the bound is tighter for larger harmonic means, and vice-versa. In particular, the norm of the weight of the non-robust and adversarially robust NTK model is large if the gradient of NTK associated with the respective model is large.

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

Definition 3. *(Bora et al. (2017)) An encoder model $\mathcal{I}$ is L-Lipschitz if $\forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$, we have:*

$$\|\mathcal{I}(\mathbf{z}_1) - \mathcal{I}(\mathbf{z}_2)\|_p \leq L\|\mathbf{z}_1 - \mathbf{z}_2\|_p. \tag{10}$$

Next, we derive the bounds of the cost difference of recourses generated for non-robust and adversarially robust models using Eqn. 10.

Theorem 3. *(Cost difference for C-CHVAE) Let $\mathbf{z}'_{\text{NR}}$ and $\mathbf{z}'_{\text{R}}$ be the solution returned by the C-CHVAE recourse method by sampling from ℓ_p-norm ball in the latent space using an L_G-Lipschitz decoder $\mathcal{G}(\cdot)$ for a non-robust and adversarially robust model. By definition of the recourse method, let $\mathbf{x}'_{\text{NR}}{=}\mathcal{G}(\mathbf{z}'_{\text{NR}}){=}\mathbf{x} + \zeta_{\text{NR}}$ and $\mathbf{x}'_{\text{R}}{=}\mathcal{G}(\mathbf{z}'_{\text{R}}){=}\mathbf{x} + \zeta_{\text{R}}$ be the respective recourses in the input space whose difference can then be bounded as:*

$$\|\zeta_{\text{NR}}\|_2 - \|\zeta_{\text{R}}\|_2 \leq \Big| \; L_G(r_{\text{R}} + r_{\text{NR}}) \; \Big|, \tag{11}$$

where 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 Def. 3, triangle inequality, L-Lipschitzness of the generative model, and the fact that the ℓ_p-norm of the model's outputs are known in the latent space. See Appendix A.4 for detailed proof. □

Implications: The right term in the Eqn. 11 entails that the ℓ_p-norm of the difference between the generated recourses using C-CHVAE is bounded if, 1) the Lipschitz constant of the decoder is small, and 2) the sum of the radii determined by C-CHVAE to successfully generate a recourse, such that they successfully return a recourse for the non-robust and adversarially robust model, is smaller.

4.2 Validity Analysis

The validity of a recourse $\mathbf{x}'$ is defined as the probability that it results in the desired outcome (Verma, Dickerson, and Hines 2020), i.e., $\Pr(f(\mathbf{x}') = 1)$. Below, we analyze the validity of the recourses generated for linear models and using Lemma 1 show that it higher for non-robust models.

Theorem 4. *(Validity comparison for linear model) For a given instance $\mathbf{x} \in \mathbb{R}^d$ and desired target label denoted by unity, let $\mathbf{x}'_{\text{R}}$ and $\mathbf{x}'_{\text{NR}}$ be the counterfactuals for adversarially robust $f_{\text{R}}(\mathbf{x})$ and non-robust $f_{\text{NR}}(\mathbf{x})$ models, respectively. Then, $\Pr(f_{\text{NR}}(\mathbf{x}'_{\text{NR}}) = 1) \geq \Pr(f_{\text{R}}(\mathbf{x}'_{\text{R}}) = 1)$ if $|f_{\text{NR}}(\mathbf{x}'_{\text{R}}) - f_{\text{NR}}(\mathbf{x}'_{\text{NR}})| \leq \Delta\|\mathbf{x}'_{\text{R}}\|_2$. Here, the validity is denoted by $\Pr(f(\mathbf{x}') = 1)$, which is the probability that the counterfactual $\mathbf{x}'$ results in the desired outcome.*

Proof Sketch. We derive the conditions under which the probability of obtaining a valid recourse is higher for a non-robust model compared to its adversarially robust counterpart using Lemma 1, natural logarithms, data processing, and Cauchy-Schwartz inequalities. See Appendix A.5 for the complete proof. □

Implications: We show that the condition for the validity is dependent on the weight difference Δ of the models (from Lemma 1). Formally, the validity of non-robust models will be greater than or equal to that of adversarially robust models only if the difference between the prediction of the non-robust model on $\mathbf{x}'_{\text{NR}}$ and $\mathbf{x}'_{\text{R}}$ is bounded by Δ times the ℓ_2-norm of $\mathbf{x}'_{\text{R}}$.

Next, we bound the weight difference of non-robust and adversarially robust wide neural networks.

Lemma 2. *(Difference between non-robust and adversarially robust weights for wide neural network models) For a given NTK model, let $\mathbf{w}^{\text{NTK}}_{\text{NR}}$ and $\mathbf{w}^{\text{NTK}}_{\text{R}}$ be weights of the non-robust and adversarially robust model. Then, for a wide neural network model with ReLU activations, the difference in the weights can be bounded as:*

$$\|\mathbf{w}^{\text{NTK}}_{\text{NR}} - \mathbf{w}^{\text{NTK}}_{\text{R}}\|_2 \leq \Delta_{\text{K}}\|\mathbf{Y}\|_2 \tag{12}$$

where $\Delta_{\text{K}}{=}\|(\mathbf{K}^\infty(\mathbf{X},\mathbf{X}){+}\beta\mathbf{I}_n)^{-1}{-}(\mathbf{K}^\infty(\mathbf{X}_{\text{R}},\mathbf{X}_{\text{R}}){+}\beta\mathbf{I}_n)^{-1}\|_2$, $\mathbf{K}^\infty$ is the kernel matrix for the NTK model defined in Def. 2, $(\mathbf{X}, \mathbf{X}_{\text{R}})$ are the training samples for the non-robust and adversarially robust NTK models, β is the bias of the ReLU NTK model, and $\mathbf{Y}$ are the labels of the training samples. Subsequently, we show that $\|\mathbf{w}^{\text{NTK}}_{\text{NR}}\|_2 - \Delta_{\text{K}}\|\mathbf{Y}\|_2 \leq \|\mathbf{w}^{\text{NTK}}_{\text{R}}\|_2 \leq \|\mathbf{w}^{\text{NTK}}_{\text{NR}}\|_2 + \Delta_{\text{K}}\|\mathbf{Y}\|_2$.

Proof Sketch. We derive the bound for the weight of the adversarially robust NTK model using the closed-form expression for the NTK weights. Using Cauchy-Schwartz and reverse triangle inequality, we prove that the ℓ_2-norm of the difference between the non-robust and adversarially robust NTK model weights is upper bounded by the difference between the kernel matrix $\mathbf{K}^\infty$ of the two models. See Appendix A.6 for detailed proof. □

Implications: Lemma 2 implies that the bound is tight if the generated adversarial samples $\mathbf{X}_{\text{R}}$ are very close to

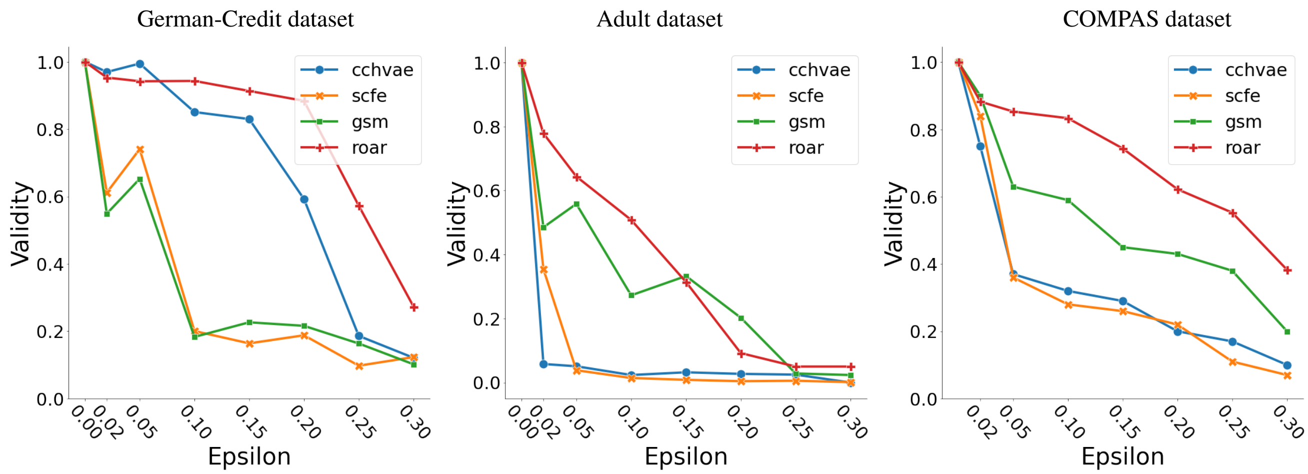


Figure 1: Analyzing validity differences between recourses generated using non-robust and adversarially robust wide neural neural networks for German Credit, Adult, and COMPAS datasets. We find that the validity decreases for increasing values of ϵ. Refer to Appendix B.1 for similar results on larger neural networks.

the original samples, i.e., the degree of robustness of the adversarially robust model is small.

Next, we show that the validity of recourses generated for non-robust wide neural network models is higher than their adversarially robust counterparts.

Theorem 5. *(Validity Comparison for wide neural network) For a given instance* $\mathbf{x} \in \mathbb{R}^d$ *and desired target label denoted by unity, let* $\mathbf{x}'_{\text{R}}$ *and* $\mathbf{x}'_{\text{NR}}$ *be the counterfactuals for adversarially robust* $f^{\text{NTK}}_{\text{R}}(\mathbf{x})$ *and non-robust* $f^{\text{NTK}}_{\text{NR}}(\mathbf{x})$ *wide neural network models respectively. Then,* $\Pr(f^{\text{NTK}}_{\text{NR}}(\mathbf{x}'_{\text{NR}}) = 1) \geq \Pr(f^{\text{NTK}}_{\text{R}}(\mathbf{x}'_{\text{R}}) = 1)$ *if* $\left\|(\mathbf{K}^{\infty}(\mathbf{x}'_{\text{R}}, \mathbf{X}_{\text{R}}) - \mathbf{K}^{\infty}(\mathbf{x}'_{\text{NR}}, \mathbf{X}))^{\text{T}}\mathbf{w}^{\text{NTK}}_{\text{NR}}\right\| \leq \left\|\mathbf{K}^{\infty}(\mathbf{x}'_{\text{R}}, \mathbf{X}_{\text{R}})^{\text{T}}\right\|\Delta_{\text{K}}\|\mathbf{Y}\|_2$. *Here, the validity is denoted by* $\Pr(f(\mathbf{x}') = 1)$*, which is the probability that the counterfactual* $\mathbf{x}'$ *results in the desired outcome.*

Proof Sketch. We extend Theorem 2 by deriving an analogous condition for wide neural network models using Lemma 2, natural logarithms, data processing, and Cauchy-Schwartz inequalities. See Appendix A.7 for the complete proof. □

Implications: Our derived conditions show that if the difference between the NTK $\mathbf{K}^{\infty}$ associated with the non-robust and adversarially robust model is bounded (i.e., $\mathbf{K}^{\infty}(\mathbf{x}'_{\text{R}}, \mathbf{X}_{\text{R}}) \approx \mathbf{K}^{\infty}(\mathbf{x}'_{\text{NR}}, \mathbf{X})$), then it is likely to have a validity greater than or equal to that of its adversarial robust counterpart. Further, we show that this bound is tighter for smaller Δ_{K}, and vice-versa.

5 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.

5.1 Experimental Setup

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

Datasets. We use three real-world datasets for our experiments: 1) The *German Credit* (Dua and Graff 2017) dataset 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* (Yeh and Lien 2009) dataset 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* (Jordan and Freiburger 2015) dataset has criminal records and demographics features for 18,876 defendants who got released on bail at the U.S state courts during 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 ReLU activation functions as our predictor and set the number of nodes in the intermediate layers

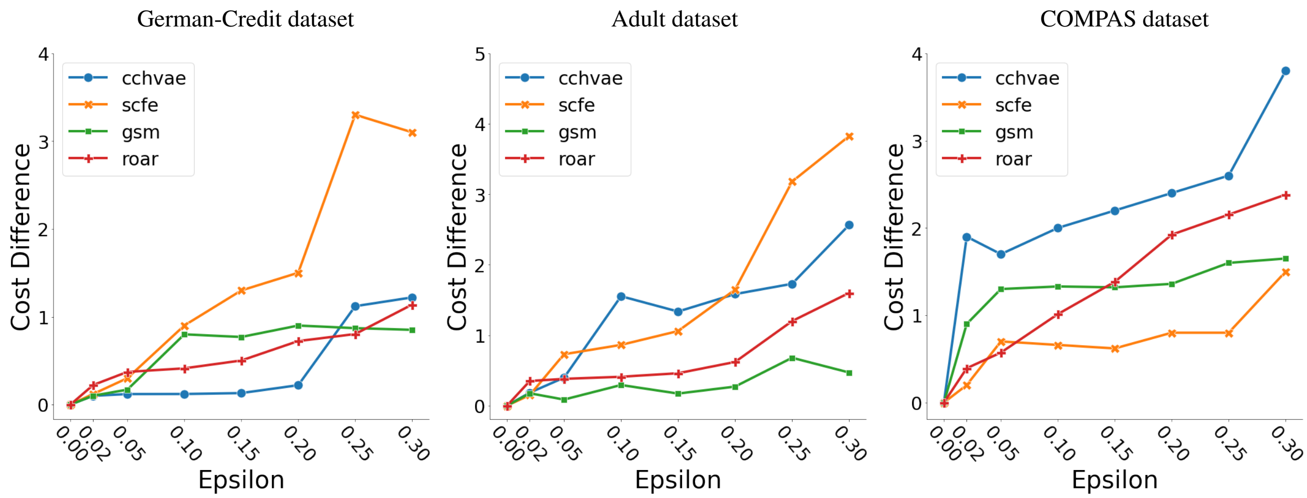


Figure 2: Analyzing cost differences between recourses generated using non-robust and adversarially robust wide neural neural networks for German Credit, Adult, and COMPAS datasets. We find that the cost difference (i.e., ℓ_2-norm) between the recourses generated for non-robust and adversarially robust models increases for increasing values of ϵ. Refer to Figure 5 for similar results on larger neural networks.

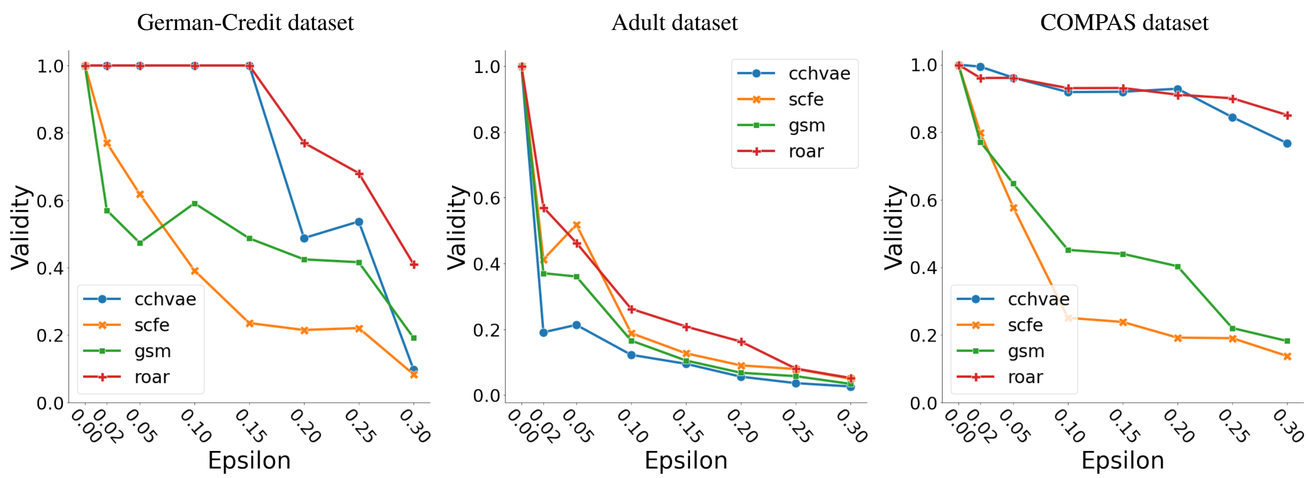


Figure 3: Analyzing validity differences between recourses generated using non-robust and adversarially robust logistic regression for German Credit, Adult, and COMPAS datasets. We find that the validity decreases for increasing values of ϵ. Refer to Appendix B.1 for similar results on larger neural networks.

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 recourses generated using four popular classes of recourse generation methods, namely, gradient-based (SCFE), manifold-based (C-CHVAE), random search-based (GSM) methods (see Sec. 3), and robust methods (ROAR) (Upadhyay, Joshi, and Lakkaraju 2021), when the underlying models are non-robust and adversarially robust.

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., $\text{Cost}(\mathbf{x}, \mathbf{x}') = \frac{1}{|\mathcal{D}_{\text{test}}|}\|\mathbf{x} - \mathbf{x}'\|_2$, where $\mathbf{x}$ is the input and $\mathbf{x}'$ is its corresponding recourse. To measure validity, we compute the probability of the generated recourse resulting in the desired outcome, i.e., $\text{Validity}(\mathbf{x}, \mathbf{x}') = \frac{|\{\mathbf{x}' : f(\mathbf{x}')=1 \,\cap\, \mathbf{x}'=g(\mathbf{x}, f)\}|}{|\mathcal{D}_{\text{test}}|}$, where $g(x, f)$ returns recourses for input $\mathbf{x}$ and predictive model f.

Implementation details. We train non-robust and adversarially robust predictive models from two popular model

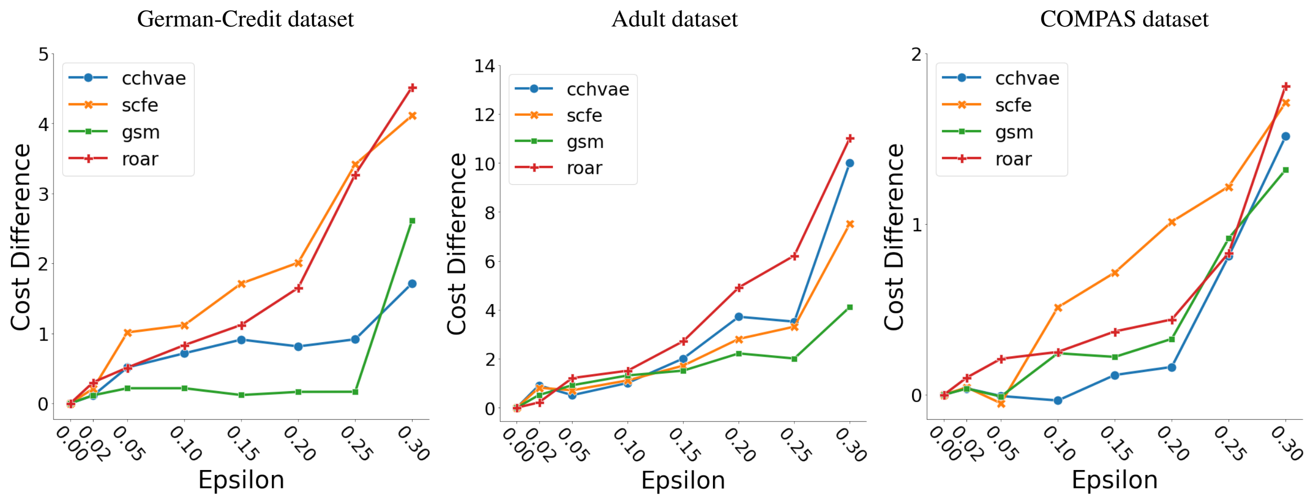


Figure 4: Analyzing cost differences between recourses generated using non-robust and adversarially robust logistic regression for German Credit, Adult, and COMPAS datasets. We find that the cost difference (i.e., ℓ_2-norm) between the recourses generated for non-robust and adversarially robust models increases for increasing values of ϵ.

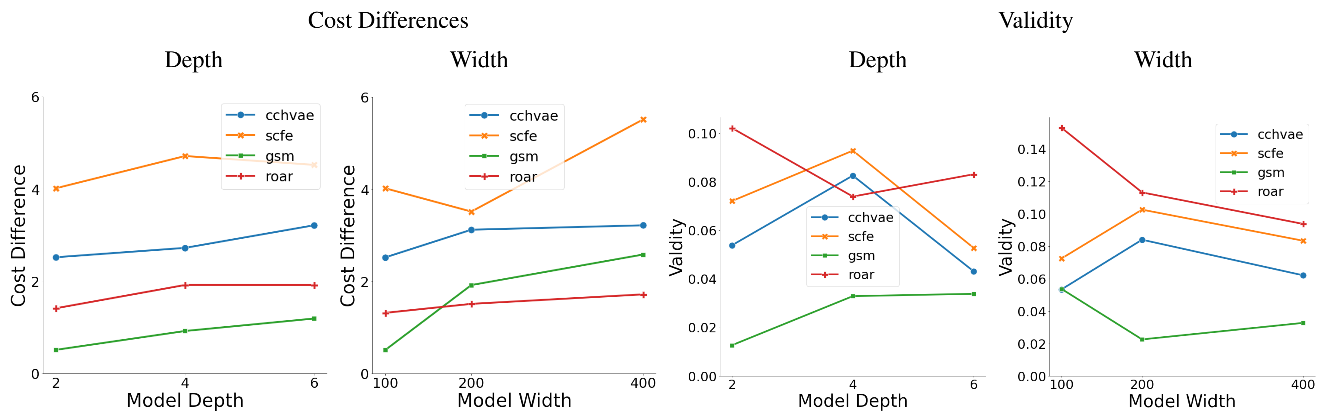


Figure 5: This figure analyzes the cost and validity differences between recourses generated using non-robust and adversarially robust neural networks trained on the Adult dataset. These differences are examined as the model size increases in terms of depth (defined as the number of hidden layers) and width (defined as the number of nodes in each hidden layer in a neural network of depth=2). Our findings suggest that: i) the cost difference (i.e., ℓ_2-norm) between the recourses generated for non-robust and adversarially robust models remains consistent even as the model's depth or width increases, and ii) the validity of the recourses remains consistent even as the model's depth or width increases. Here, the adversarially robust model is trained with $\epsilon = 0.3$.

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 ϵ=0 is the non-robust model.

5.2 Empirical Analysis

Next, we describe the experiments we carried out to understand the impact of adversarial robustness of predictive models on algorithmic recourse. More specifically, we discuss the (1) empirical verification of our theoretical bounds, (2) empirical analysis of the differences between the costs of recourses when the underlying model is non-robust vs. adversarially robust, 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 4 on real-world datasets. In particular, we first estimate the empirical bounds (RHS of Theorems 1,3) 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-based and manifold-based recourse methods

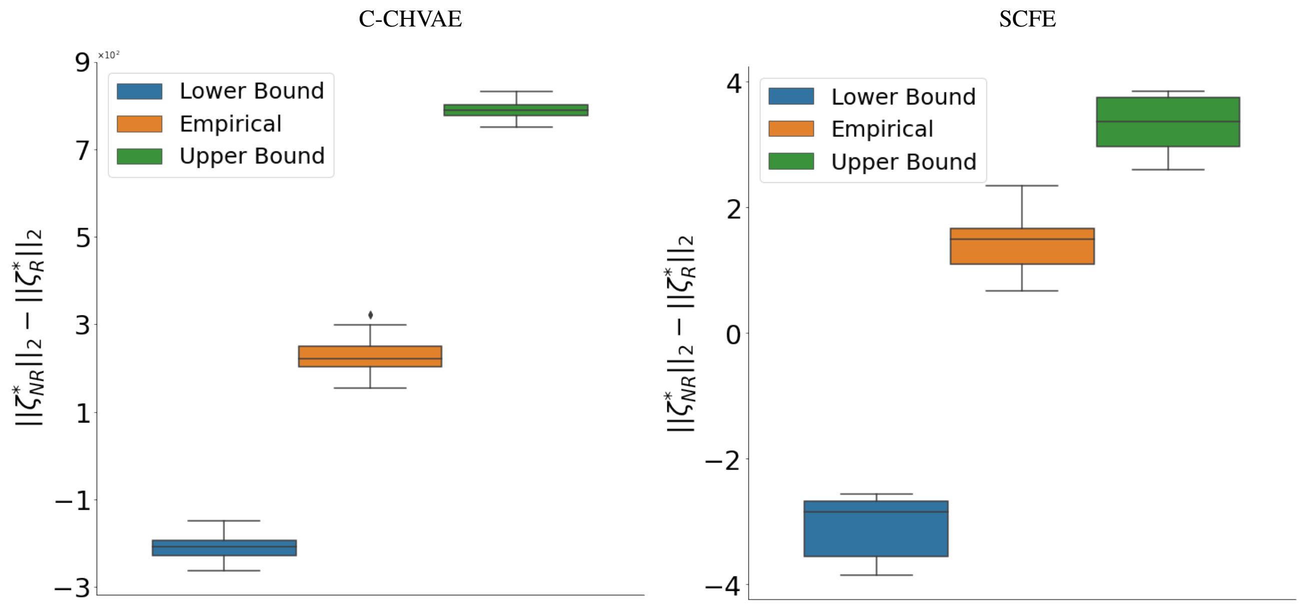
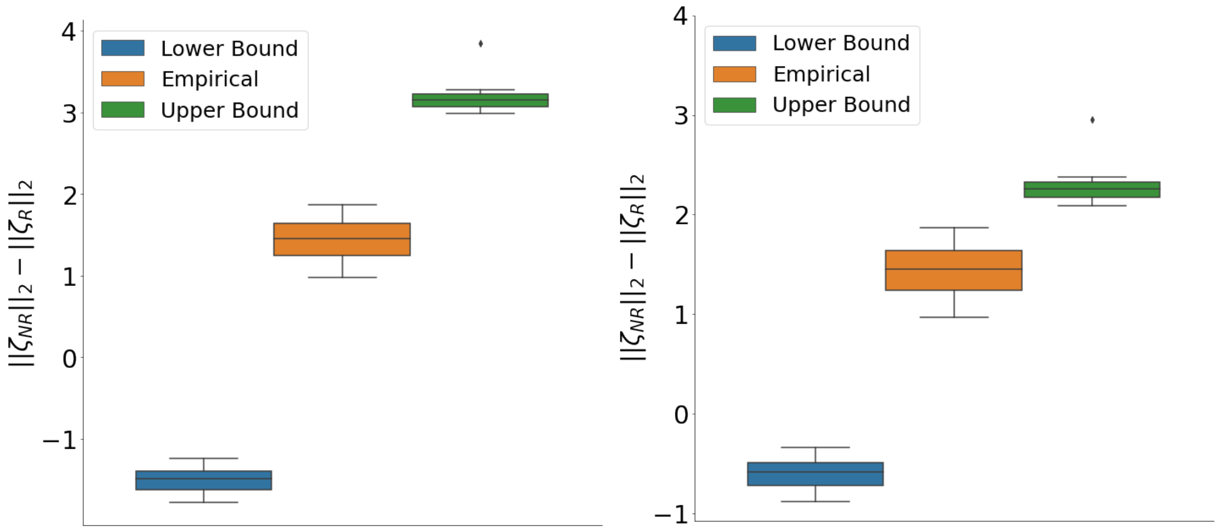


Figure 6: Cost differences (left): Empirically calculated cost differences (in orange) and our theoretical lower (in blue) and upper (in green) bounds for C-CHVAE and SCFE recourses corresponding to adversarially robust (trained using ϵ=0.3) vs. non-robust linear models trained on the Adult dataset. See Fig. 7 for similar bounds for adversarially robust (ϵ=0.3) vs. non-robust neural-network model. Validity (right): Empirical difference between the validity of recourses for non-robust and adversarially robust linear and neural network model trained on Adult dataset. Results show no violations of our theoretical bounds.



(a) (b)

Figure 7: (a) Empirically calculated cost differences (in orange) for the original model and our theoretical lower (in blue) and upper (in green) bounds for SCFE recourses corresponding to adversarially robust (trained using ϵ=0.1) vs. non-robust neural networks corresponding to test samples of the Adult dataset, based on Theorem 2. (b) Empirically calculated cost differences (in orange) for the original model and our theoretical lower (in blue) and upper (in green) bounds for SCFE recourses corresponding to adversarially robust (trained using ϵ=0.2) vs. non-robust neural networks corresponding to test samples of the Adult dataset, based on Theorem 2.

(LHS of Theorems 1,3). Figure 6 shows 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. A similar trend is

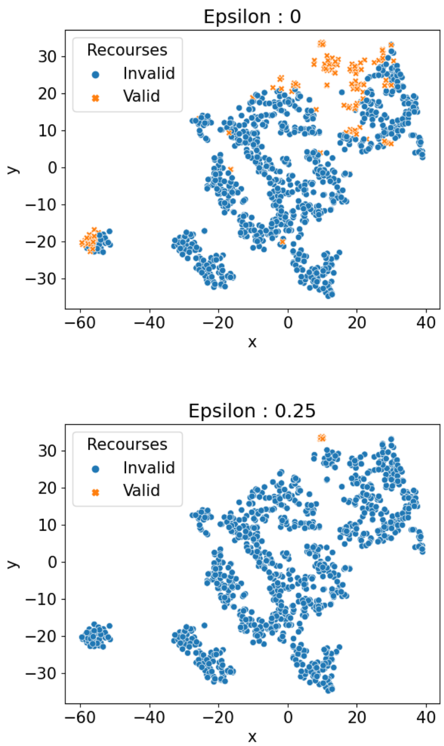


Figure 8: A t-SNE visualization of the change in availability of valid recourses (orange) for adversarially robust models trained using $\epsilon = [0, 0.25]$, where a non-robust model is a model trained using $\epsilon = 0$. Results are shown for a neural network model trained on the Adult dataset. We observe fewer valid recourses for higher values of ϵ in this local neighborhood.

observed for Theorem 2 which is for the case of non-linear models, shown in Figure 7. For the case of theoretical bounds for validity analysis in Theorem 4, we observe that the validity of the non-robust model (denoted by $\Pr(f_{\text{NR}}(x) = 1)$ in Theorem 4) was higher than the validity of the adversarially robust model for all the test samples following the condition in Theorem 4 ($> 90\%$ samples) for a large number of training iterations used for training adversarially robust models with $\epsilon \in \{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$, shown Figure 6.

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 ϵ. Results in Figure 4 show a significant increase in costs to find algorithmic recourse for adversarially robust neural network models with increasing degrees of robustness for all the datasets. In addition, the recourse cost for adversarially robust model is always higher than that of non-robust model (see appendix Figure 10 for similar trends for logistic regression model). Further, we observe a relatively smoother increasing trend for SCFE cost differences compared to others. We attribute this trend to the stochasticity present in C-CHVAE and GSM. We find a higher cost difference in SCFE for most datasets, which could result from the larger sample size used in C-CHVAE and GSM. Further, we observe a similar trend in cost differences on increasing the number of iterations to find a 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 a non-robust and adversarially robust model under resource constraints, and plot it against varying degrees of robustness ϵ. Results in Figure 6 show a strong impact of adversarial training on the validity for logistic regression and neural network models trained on three real-world datasets (also see Appendix B). On average, we observe that the validity drops to zero for models adversarially trained with $\epsilon > 0.2$. To shed more light on this observation, we use t-SNE visualization (Van der Maaten and Hinton 2008) – a non-linear dimensionality reduction technique – to visualize test samples in the dataset to two-dimensional space. In Figure 8, we observe a gradual decline in the number of valid recourses around a local neighborhood with an increasing degree of robustness ϵ. The decline in the number of valid recourses suggests that multiple recourses in the neighborhood of the input sample are classified to the same class as the input for higher ϵ, supporting our hypothesis that adversarially robust models severely impact the validity of recourses and make the recourse search computationally expensive.

6 Conclusion

In this work, we theoretically and empirically analyzed the impact of adversarially robust models on actionable explanations. We theoretically bounded the cost differences between recourses output by state-of-the-art techniques when the underlying models are adversarially robust vs. non-robust. We also bounded the validity differences between the recourses corresponding to adversarially robust vs. non-robust models. Further, we empirically validated our theoretical results using three real-world datasets and two popular classes of predictive models. Our theoretical and empirical analyses demonstrate that adversarially robust models significantly increase the cost and decrease the validity of the resulting recourses, thereby highlighting the inherent trade-offs between achieving adversarial robustness in predictive models and providing 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 that 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 to secure lower cost recourses.

References

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

Bansal, N.; Agarwal, C.; and Nguyen, A. 2020. Sam: The sensitivity of attribution methods to hyperparameters. In *CVPR*.

Bora, A.; Jalal, A.; Price, E.; and Dimakis, A. G. 2017. Compressed sensing using generative models. In *ICML*. PMLR.

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

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

Dominguez-Olmedo, R.; Karimi, A.-H.; and Schölkopf, B. 2021. On the Adversarial Robustness of Causal Algorithmic Recourse. *arXiv:2112.11313*.

Du, S. S.; Zhai, X.; Poczos, B.; and Singh, A. 2019. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *ICLR*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Freiesleben, T. 2022. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32(1): 77–109.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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

Jordan, K. L.; and Freiburger, T. L. 2015. The Effect of Race/Ethnicity on Sentencing: Examining Sentence Type, Jail Length, and Prison Length. *Journal of Ethnicity in Criminal Justice*, 13(3): 179–196.

Karimi, A.-H.; Barthe, G.; Balle, B.; and Valera, I. 2020a. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

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

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

Kolter, Z.; and Madry, A. 2023. Adversarial Robustness - Theory and Practice. https://adversarial-ml-tutorial.org/linear_models/.

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

Laugel, T.; Lesot, M.-J.; Marsala, C.; Renard, X.; and Detyniecki, M. 2017. Inverse Classification for Comparison-based Interpretability in Machine Learning. *arXiv preprint arXiv:1712.08443*.

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

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

Pawelczyk, M.; Agarwal, C.; Joshi, S.; Upadhyay, S.; and Lakkaraju, H. 2022a. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In *AISTATS*.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*.

Pawelczyk, M.; Datta, T.; van-den Heuvel, J.; Kasneci, G.; and Lakkaraju, H. 2022b. Algorithmic Recourse in the Face of Noisy Human Responses. *arXiv preprint arXiv:2203.06768*.

Rawal, K.; Kamar, E.; and Lakkaraju, H. 2021. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. *arXiv:2012.11788*.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *NeurIPS*.

Shah, H.; Jain, P.; and Netrapalli, P. 2021. Do Input Gradients Highlight Discriminative Features? *Advances in Neural Information Processing Systems*, 34: 2046–2059.

Slack, D.; Hilgard, S.; Lakkaraju, H.; and Singh, S. 2021. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

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

Upadhyay, S.; Joshi, S.; and Lakkaraju, H. 2021. Towards Robust and Reliable Algorithmic Recourse. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Van Looveren, A.; and Klaise, J. 2019. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*.

Venkatasubramanian, S.; and Alfano, M. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. New York, NY, USA: ACM.

Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual Explanations for Machine Learning: A Review. *arXiv:2010.10596*.

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

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2).

Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A. L.; and Le, Q. V. 2020. Adversarial examples improve image recognition. In *CVPR*.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. In *Expert Systems with Applications*.

Zhang, R.; and Zhang, S. 2022. Rethinking Influence Functions of Neural Networks in the Over-Parameterized Regime. In *AAAI*.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.