UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Towards Improved Recourse Trade-offs with Adaptive Adversarial Training

A thesis submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE & ENGINEERING
by

Ian Hardy

March 2023

The Thesis of Ian Hardy is approved:

Professor Yang Liu, Chair

Professor Leilani Gilpin

Professor Lise Getoor

Peter Biehl
Vice Provost & Dean of Graduate Studies

Contents

List of Figures

Towards Improved Recourse Trade-offs with Adaptive Adversarial Training

Ian Hardy

Abstract

Recent work has connected adversarial attack methods and algorithmic recourse methods: both seek minimal changes to an input instance which alter a model's classification decision. It has been shown that traditional adversarial training, which seeks to minimize a classifier's susceptibility to malicious perturbations, increases the cost of generated recourse; with larger attack tolerances (known as attack radii) during adversarial training correlating with higher recourse costs. From the perspective of algorithmic recourse, however, the appropriate adversarial training radius has always been unknown. Another recent line of work has motivated adversarial training with adaptive adversarial training radii to address the issue of instance-wise variable adversarial vulnerability, showing success in domains with unknown attack radii. This work studies the effects of adaptive adversarial training on algorithmic recourse costs, establishing that the improvements in model robustness induced by adaptive adversarial training show minimal effects on algorithmic recourse costs. This provides a potential avenue for affordable robustness in domains where recoursability is critical.

Acknowledgments

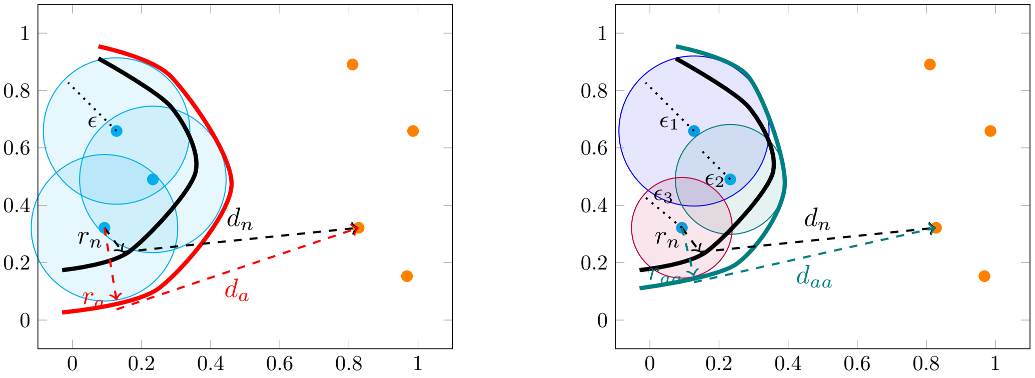
I would like to thank Professor Yang Liu for his guidance throughout this work, it has been an honor to work with you. Special thanks to Jayanth Yetukuri for his advice, suggestions, and help. They were invaluable.

1 Introduction

The adoption of Machine Learning (ML) in consequential environments motivates the provision of instructions to adversely-affected users on actions they can take to alter a model's decision. For example, in the lending domain, if a classifier decides to deny an applicant, there should be a mechanism for providing a feasible set of actions the applicant can take to be approved. This instructive information is desirable as opaque self-learning systems inform more and more of our society's decision-making, for both trust and accountability. The ability to obtain a desired outcome from a known model, the actionable set of changes that users can make to improve their qualification, or the systematic process of reversing unfavorable decisions is defined as "algorithmic recourse," or simply "recourse" [11]. These what-if scenarios are also often referred to as "counterfactual explanations." Importantly, the explicitly stated goal of recourse is to find actions with minimal cost to the user [23].

Simultaneously, it has been observed that many neural networks can be easily "fooled" by introducing small changes to input features that may seem imperceptible. [22] first proposed the concept of "adversarial examples": by adding small perturbations to an input sample, models obtain incorrect classification results with high confidence scores. These are sometimes referred to as "evasion attacks" [5]. [22] also found that such perturbations can be adapted into different model architectures, demonstrating that many deep neural networks are vulnerable to these input manipulations. Adversarial examples raise concerns about the trust one can place in neural network classifiers, and much work has been put into adversarial training methods to improve the robustness of models to adversarial examples. The most popular adversarial training regimes [1] generate adversarial examples (with corrected labels) within a fixed "attack radius" (ϵ) during training procedure and include them in the model's training dataset. While adversarial training has been shown to increase robustness to adversarial examples drastically, it often comes at some cost to standard accuracy [27].

There is an inherent contention between the considerations of algorithmic recourse and adversarial robustness. While minimizing the changes necessary to alter a classifier's decision is seen as beneficial from a recourse perspective, such changes are harmful from a robustness perspective. Research [14] has demonstrated that adversarial training increases the average recourse cost, with higher adversarial training radii corresponding to higher recourse costs, which raises



(a) Toy problem demonstrating that adversarial training can result in counterfactuals that are both costlier and further from the desired class manifold. The natural decision boundary is shown in black, the adversarial boundary in red. ϵ-Adversarial training creates a necessary recourse cost $c_a = \epsilon > c_n$, and yields a distance in the resulting recourse to the desired manifold of $d_a > d_n$

(b) Adaptive Adversarial Training provides counterfactuals which are cheaper and relatively closer to the desired class manifold. The natural decision boundary is shown in black, the adaptive adversarial boundary in green. With instance specific robustness ϵ_i, the recourse cost $c_{aa} = \epsilon_i > c_n$ and $c_{aa} < \epsilon$ for any $\epsilon_i < \epsilon$. This yields a distance $d_{aa} < d_a$.

Figure 1: An example scenario demonstrating the effectiveness of AAT in terms of recourse costs.

the concern that there may be an inherent trade-off between robustness and recourse.

Briefly, it should be noted that the goals of adversarial robustness are not totally at odds with recourse. Recourse *should* represent *true movements towards a desired class*, and adversarial examples that "fool" a model can be *harmful* and should *not be presented as recourse*. Consider the lending setting: if an approval action plan is provided to an applicant which does not represent a true movement in their underlying propensity for repayment, both the lender and borrower are putting themselves at long-term financial risk by following that plan. This is relevant in the context of many recourse settings, where data is tabular and it is not immediately obvious which input perturbations constitute adversarial examples and which input perturbations constitute recourse that genuinely moves an individual towards a desired class manifold. With this in mind, it is worth considering not only the change in overall cost of recourse, but also the change in proximity of recourse to the desired data manifold, when selecting an adversarial training radius.

Even more fundamentally, it is important to question *whether a fixed adversarial training radius is appropriate, particularly in the context of algorithmic recourse?* It has been shown [2] that different data instances have different *inherent adversarial vulnerabilities* due to their varying proximities to other classes. As such, some researchers have argued that an identical adversarial training radius should not be applied to all instances during training. Several methods [2, 6, 8] have been proposed for automatically learning *instance-wise* adversarial radii to address this variability. These are broadly referred to as "Adaptive Adversarial Training" (AAT) regimes [1].

This work explores the effects of AAT on both model robustness and ultimate recourse costs in an attempt to address the trade-off between the two and find a *justifiable* middle ground. The contributions include:

- An observation on the effects of robustness on recourse costs, and when AAT yields more affordable recourse.
- Experiments demonstrating AAT's superior robustness/ recourse trade-offs over traditional AT.

2 Background and Related works

Algorithmic Recourse: The continued adoption of ML in high-impact decision making such as banking, healthcare, and resource allocation has inspired much work in the field of Algorithmic Recourse [23, 13, 12], and Counterfactual Explanations [19, 26, 15, 21]. The performance of different recourse methods depends highly on properties of the datasets they are applied to, the model they operate on, the application of that model's score, and factual point specificities [7]. However, broadly speaking, recourse methods are classified based on: i) the *model family* they apply to, ii) the degree of *access* they have to the underlying model (i.e. white vs. black box methods), iii) the consideration of *manifold proximity* in the generation of recourse, iv) the underlying *causal relationships* in the data, and v) the use of *model approximations* in the generation process [25]. Recently, [18] introduced CARLA, a framework for benchmarking different recourse methods which act as an aggregator for popular recourse methods and standard datasets.

Adversarial Attacks and Adversarial Training: Adversarial vulnerability refers to the susceptibility of a model to be *fooled* by perturbations to the input data which cannot be detected by humans (so-called *Adversarial Examples*) [22]. Adversarial Training [10, 16] has been introduced to create models which are not susceptible to such attacks. The most popular method of Adversarial Training generates adversarial examples during the training process and includes them in the training dataset with corrected labels alongside the uncorrupted dataset. Often, adversarial training comes at some cost to standard classification accuracy. There have been many attack methods proposed to generate adversarial examples [5] with varying degrees of access to the model under attack, but most focus on defending against adversarial examples within a given ϵ-radius (which are often defined by ℓ_1, ℓ_2, or ℓ_∞ norms of size ϵ.) This work follows the popular attack and training formulation from [16], which minimizes the worst-case loss within a defined ϵ-radius.

On the Intersection of Robustness and Recourse Both Adversarial Examples and Counterfactual Explanations are formally described as constrained optimization problems where the objective is to alter a model's output by minimally perturbing input features [4, 9]. Recent work [17] proved equivalence between certain adversarial attack methods and counterfactual explanation methods, and further work has demonstrated both theoretically and empirically that increasing the radius of attack during adversarial training increases the cost of the resulting recourse [14]. This inherent connection pits security at odds with expressivity and raises an important question as to how an adversarial radius ought to be selected for adversarial training. If the radius is too small, the model may be overly sensitive to an attack, while if it is too large, end users suffer from potentially overly-burdensome recourse costs. In the context of many recourse problems where data is tabular, it is difficult to determine what may constitute an adversarial attack, furthering the difficulty of radius selection. [3] discussed a formulation for adversarial attacks on tabular data that accounts for both the radius of attack and the importance of a feature, but this is difficult to know a priori and often changes depending on the choice of explanation method selected [20].

Adaptive Adversarial Training It has been observed that different data instances have different inherent adversarial vulnerability due to their varying proximity to other class' data manifolds, calling into question the conventional wisdom that models should be adversarially trained at a single consistent adversarial radius. [2] first observed this issue in the image classification domain, where certain instances can be *meaningfully transformed* into other classes even at small adversarial radii. The authors of [2] proposed a means of discovering instance-wise adversarial radii by iteratively increasing or decreasing each instance's attack radius based on whether attacks are successful. [6] built on this work by further motivating the effects of overly-large adversarial radii on classification accuracy and proposed a variation of [2]'s method which included adaptive label-smoothing to account for the uncertainty added by larger attack radii, and [8] proposed a means for adaptive adversarial training by increasing the classification margin around correctly-classified datapoints. Adaptive Adversarial Training (AAT) presents a means of "automatically" selecting attack radii during training, and in all works thus far, has shown positive results in terms of the accuracy/robustness trade-off inherent in adversarial training, as well as smoother robustness curves across ranges of attack radii compared with traditional Adversarial Training.

3 Preliminaries & Notation

Standard Training: We begin with a model f parameterized by weights θ that maps $\mathcal{X} \to \mathcal{Y}$, where $x \in \mathcal{X}$ are features and $y \in \mathcal{Y}$ are their corresponding labels. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, and a loss function $\ell(\cdot)$, a standard learning objective is to minimize the average loss on the data:

$$\min_{\theta} \quad \frac{1}{N} \sum_{(x_i, y_i) \in D} \ell(f_\theta(x_i), y_i) \tag{1}$$

Let f_{nat} represent the naturally trained model using the standard loss minimization based optimization technique.

Adversarial Attacks: The goal of an adversarial attack is to strategically generate perturbations δ which can significantly enlarge the loss $\ell(\cdot)$ when added to an instance x. [10] introduced *Fast Gradient Sign Method (FGSM)* for generating adversarial examples using the following mechanism:

$$x_i' = x_i + \alpha \cdot \operatorname{sign}\left(\nabla_{x_i} \ell\left(f_\theta(x_i), y_i\right)\right) \tag{2}$$

where α denotes the size of the perturbation, x_i' denotes the adversarially perturbed sample, and x_i is the original clean sample. The *sign* function operates on the gradient of $\ell\left(f_\theta(x_i), y_i\right))$

w.r.t. x_i, which is used to set the gradient to 1 if it is greater than 0 and -1 if it is less than 0. [16] proposed a stronger iterative version of FGSM, performing Projected Gradient Descent (PGD) on the negative loss function:

$$x_i(t+1) = \Pi_{x+\mathcal{S}} \left(x_i(t) + \alpha \cdot \operatorname{sign} \left(\nabla_{x_i(t)} \ell(f_\theta(x_i(t)), y_i) \right) \right)$$

where α denotes the perturbation step size at each iteration and $x_i(t{+}1)$ represents the perturbed example at step $t+1$ for the clean instance x_i. In this work, I use PGD due to its performance, popularity, and relative speed.

Adversarial Training: Adversarial training is usually formulated as a min-max learning objective, wherein we seek to minimize the worst case loss within a fixed training radius ϵ.

$$\min_\theta \max_{||\delta_i|| \le \epsilon} \frac{1}{N} \sum_{(x_i, y_i) \in D} \ell(f_\theta(x_i + \delta_i), y_i) \tag{3}$$

We solve this min-max objective via an alternating stochastic method that takes minimization steps for θ, followed by maximization steps that approximately solve the inner optimization using k steps of an adversarial attack. PGD with a fixed ϵ is used to perturb an original instance and let $f_{\epsilon\text{-adv}}$ represent the model trained with a PGD radius of ϵ.

3.1 Adaptive Adversarial Training

[2] first argued that different data instances have different intrinsic adversarial vulnerabilities due to their varying proximity to other class manifolds, and introduced Instance-Adaptive Adversarial Training (AAT) to automatically learn instance-wise adversarial radii. The authors proposed the following objective function:

$$\min_\theta \max_{||\delta_i|| \le \epsilon_i} \frac{1}{N} \sum_{(x_i, y_i) \in D} \ell(f_\theta(x_i + \delta_i), y_i) \tag{4}$$

where ϵ_i denotes each training instance's attack radius. ϵ_i is iteratively updated at each training epoch, increasing by a constant factor if the attack at the existing radius is unsuccessful and decreasing by a constant factor if it is successful.

[8] presented an alternate form of AAT called Max-Margin Adversarial (MMA) Training that

seeks to impart adversarial robustness by maximizing the margin between correctly classified datapoints and the model's decision boundary. Formally, they proposed the following objective:

$$\min_{\theta} \left\{ \sum_{i \in \mathcal{S}_{\theta}^{+}} \max\{0, d_{max} - d_{\theta}(x_i, y_i)\} + \beta \sum_{i \in \mathcal{S}_{\theta}^{-}} \ell(f_{\theta}(x_j), y_j) \right\} \tag{5}$$

where S_{θ}^{+} is the set of correctly classified examples, S_{θ}^{-} is the set of incorrectly classified examples, $d_{\theta}(x_i, y_i)$ is the margin between correctly classified examples and the model's decision boundary, d_{max} is a hyper-parameter controlling which points to maximize the boundary around (forcing the learning to focus on points with d_{θ} less than d_{max},) and β is a term controlling the trade-off between standard loss and *margin maximization*. The authors use a line search based on PGD to efficiently approximate $d_{\theta}(x_i, y_i)$. For the rest of this study, let f_{aat} be a model trained using a mechanism from this category of training techniques.

3.2 Recourse Methods

For the scope of this study, I explore three different classes [14] of recourse methods: i) one random search, ii) one gradient-based search, and iii) one manifold-based approach. I will now briefly discuss each method, and refer the readers to the original works for further implementation details.

Growing Spheres (GS): [15] proposed a random search method for calculating counterfactual by sampling from points within ℓ_2-hyper-spheres around x of iteratively increasing radii until one or more counterfactual is identified which flips $f(x)$. Formally, they present a minimization problem in selecting which counterfactual x' to return:

$$\arg\min_{x' \in \mathcal{X}} \{c(x, x') | f(x) \neq f(x')\} \tag{6}$$

where $\mathcal{X}$ is the family of sampled points around x and c is a cost function in $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{+}$: $||x' - x||_2 + \gamma ||x' - x||_0$, where γ is a hyperparameter controlling the desired sparsity of the resulting counterfactual.

Score Counterfactual Explanations (SCFE): [26] proposed a gradient-based method for identifying counterfactuals x'.

$$\arg\min_{x'}\max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{7}$$

where $d(\cdot, \cdot)$ is some distance function and y' is the desired score from the model. In practice, this is solved by iteratively finding x' and increasing λ until a satisfactory solution is identified.

CCHVAE: [19] proposed a manifold-based solution to finding counterfactuals using a Variational Auto Encoder (VAE) to search for counterfactuals in a latent representation $\mathcal{Z}$. The goal of CCHVAE and other manifold methods is to find counterfactuals that are semantically "similar" to other data points. Formally, given an encoder $\mathcal{E}$, a decoder $\mathcal{H}$, and a latent representation $\mathcal{Z}$ where $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$, CCHVAE optimizes the following:

$$\arg\min_{z' \in \mathcal{Z}} \{||z'|| \ s.t. \ f(\mathcal{H}(\mathcal{E}(x) + z')) \neq f(x)\} \tag{8}$$

4 Recourse trade-offs with Adaptive Adversarial training

Recourse cost The cost of recourse is usually approximated using a distance based metric. A common practice among recourse methodologies is to minimize the cost in some form or the other, because in general a low cost recourse is assumed to be easier to act upon. The cost of a recourse for a classification based model is traditionally interpreted as the minimum distance between a factual and the decision boundary. Alternatively, the inherent goal of adversarial training is to maximize the distance between factuals and the decision boundary. Hence, traditional adversarial training exacerbates the recourse costs of a classifier. In this section, I make preliminary observations on the effects of adaptive adversarial training on recourse costs.

An increase in ϵ for ϵ-adversarial training increases the overall recourse costs and the corresponding relation between ϵ and C is discussed in [14]. In comparison with an ϵ-adversarial training, I observe the following benefits from the instance adaptive adversarially training:

4.1 Recourse Costs

Let $\delta_x^{(nat)} = d(x, x')$ be the distance to the closest adversarial example x' for the instance x for a standard training based model, and, analogously, let $c_x^{(nat)} = cost(x, x'')$ be the cost of a recourse x'' for an individual represented by x. For simplicity, I assume that both $c_{(\cdot)}^{(\cdot)}$ and $cost(\cdot, \cdot)$ use the same ℓ_p norm based distance metrics. Let $H^- = \{x \in \mathcal{X} : f(x) = -1\}$ represent the sub-population which was adversely affected by the classifier $f(\cdot)$, and analogously let $H^+ = \{x \in \mathcal{X} : f(x) = +1\}$ represent the population that is favorably affected. The average cost of recourses for H^- is defined for a naturally trained model as:

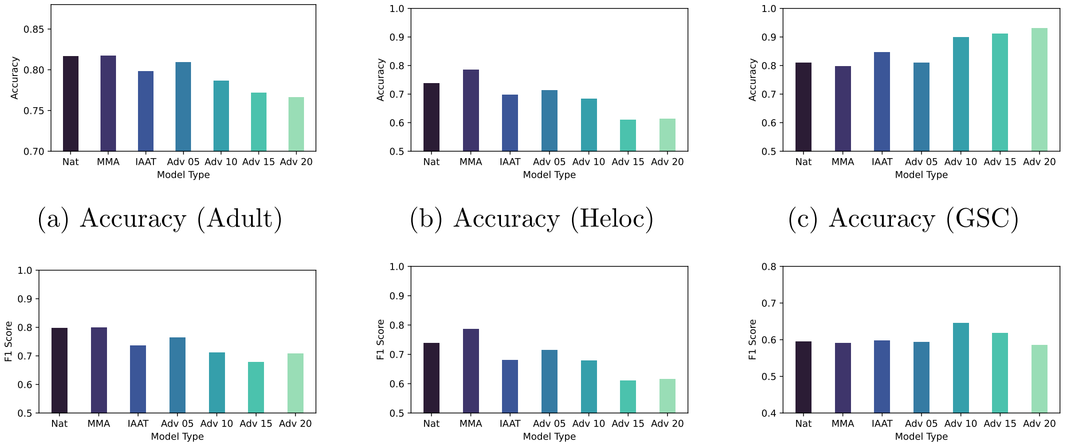
$$c_*^{(nat)} = \frac{1}{|H^-|} \sum_{x \in H^-} c_x^{(nat)} \tag{9}$$

Let $\underline{H}^- = \{x \in \mathcal{X} : f(x) = -1, c_x^{(nat)} \leq \underline{\epsilon}\}$, where $\underline{\epsilon}$ is a cost threshold to identify low cost recourses. As observed in Figures 4 and 5, a low cost counterfactual is sufficient in practice for a large section of the population. However, an optimal ϵ_a-adv classifier provides at least ϵ_a robustness to all samples in the training dataset. This can be visualized by the sharp peak in the distribution of the observed ϵ in the test dataset for all the ϵ-adv models (Figure 8). However AAT models provide natural robustness to the data samples, meaning that a data instance closer to the natural decision boundary has $\epsilon_{aat}^{H^-}$ that depends on the data's natural proximity to the decision boundary. For instances with $\epsilon_{aat}^{H^-} < \epsilon_a$, the resulting recourse will be more affordable. For $\epsilon_{aat}^{H^-} < c_x^{(nat)}$, low cost recourse within $\underline{H}^-$ will be preserved.

4.2 Proximity to the Desired Manifold

Manifold Proximity measures the distance by some metric between recourse and the target sub-population. For an $f^*_{\epsilon_a\text{-adv}}$ model, the recourse suggested have at least ϵ_a proximity from the target approved sub-population H^+ due to the fact that the target sub-population is also ϵ_a away from the decision boundary. Alternatively f_{aat} is naturally robust for the target sub-population as well. Hence, the Recourse provided has the potential to be closer in terms of proximity to H^+, so long as $\epsilon_{aat}^{H^+} < \epsilon_a$. I report the average proximity $\rho_{f_{\epsilon\text{-adv}}}$ of the model $f_{\epsilon\text{-adv}}$ using:

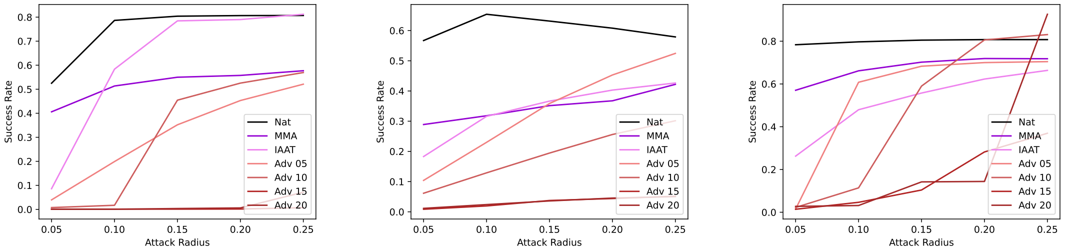
$$\rho_{f_{\epsilon\text{-adv}}} = \frac{1}{|N_{test}|} \sum_{x \in N_{test}} \min_{x^+ \in H^+} d(x, x^+) \tag{10}$$



(a) Accuracy (Adult) (b) Accuracy (Heloc) (c) Accuracy (GSC)

(d) F1 (Adult) (e) F1 (Heloc) (f) F1 (GSC)

Figure 2: Standard performance across datatsets. MMA shows particularly competitive standard performance compared with all other Adversarial Training regimens.



(a) Adult (b) Heloc (c) GSC

Figure 3: Attack Success Rate. Traditional Adversarial Training shows higher robustness within its predefined training threshold, but sharper robustness degradation as the attack radius increases.

where $d(x, x^+)$ is a distance measure between a counterfactual x and a target population x^+. I report both $\rho_{f_{\epsilon\text{-adv}}}$ and $\rho_{f_{aat}}$ for the corresponding models. In Figure 7, we see that $\rho_{f_{aat}}$ is significantly better than $\rho_{f_{\epsilon\text{-adv}}}$. A motivating toy problem demonstrating lower recourse costs and closer manifold proximity is also visualized in Figure 1.

4.3 Preservation of Low Cost Recourse

The recourse costs provided to the adversely affected individuals by a model should follow the natural distribution of the difficulty of acting upon the suggested recourse at the population level. With a fixed ϵ while training an optimal adversarially trained $f^*_{\epsilon\text{-adv}}$ model, the recourse suggested must necessarily be ϵ away from the decision boundary and further ϵ away from the nearest target population sample. Such counterfactuals contradicts with the recourse literature [24], which describes a distribution in recourse costs wherein a proportion of individuals only require minimal low cost actionable steps to obtain the desired outcome from a model, whereas other individuals can have a much larger recourse costs. Essentially, ϵ-robustness necessarily denies recourse with lower costs than ϵ.

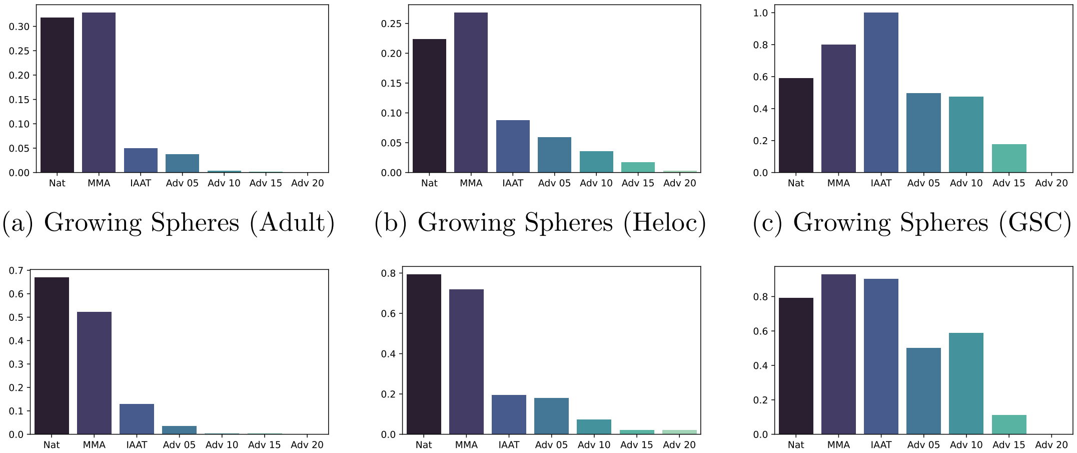
f_{aat} does not enforce a strict ϵ while training, allowing instances to have a wider range of recourse costs. To this end I compare the rate of extreme low cost recourse C_Δ across the discussed training methods with real-world datasets to measure the rate at which it degrades in practice. For simplicity, I measure:

$$C_\Delta = \frac{1}{|N_{test}|} \sum_{x_i \in N_{test}} \mathbf{1}(C_{x_i} < \epsilon) \tag{11}$$

where C_{x_i} is the cost of recourse for an instance x_i and ϵ is a minimum adversarial training radius. We can observe in Figure 4 that Adaptive Adversarial Training preserves low cost recourse rates despite providing overall robustness benefits.

5 Experimental Design & Metrics

In this section, I detail my experimentation procedure to empirically evaluate these various training methods and explain my metric choices. The CARLA package [18] was used to source the datasets and recourse methods employed.



(a) Growing Spheres (Adult)

(b) Growing Spheres (Heloc)

(c) Growing Spheres (GSC)

(d) SCFE (Adult)

(e) SCFE (Heloc)

(f) SCFE (GSC)

Figure 4: Low cost recourse (ℓ_∞ ¡ 0.05) proportion for methods that optimize directly in the input space. We can observe that AAT models has much higher proportions of low cost recourse, supporting the hypothesis that it allows for robustness while preserving low recourse costs for individuals near natural decision boundaries.

5.1 Experimental Setup

Datasets I performed my experiments on three datasets:

- *Adult Income*: A dataset originating from the 1994 Census of 48,842 individuals for whom the task is to predict whether someone makes more than $50,000/yr. It is comprised of 20 features which are a combination of demographic features (age, sex, racial group), as well as employment features (hours of work per week and salary), and financial features (capital gains/losses.) In keeping with [14] and [3], I removed categorical features for efficient training and approximation of tabular adversarial examples. The target distribution is somewhat skewed, with a 76% positive label proportion.

- *Home Equity Line of Credit (Heloc)*: pulled from the 2019 FICO Explainable Machine Learning (xML) challenge, the Heloc dataset consists of anonymized credit bureau data from 9,871 individuals where the task is to predict whether an individual will repay their HELOC account within two years. The dataset consists of 21 financial features and no demographic data. The target distribution is evenly split, with a 48% positive label proportion.

- *Give Me Some Credit (GSC)*: a credit-scoring dataset pulled from a 2011 Kaggle Competition consisting of 150,000 individuals for whom the task is to predict default. It consists of 11 features, one of which is a demographic feature (age), and the rest are financial variables. The target distribution is heavily skewed, with a 93% positive label proportion.

Models I trained a total of 7 Neural Network models for each of the datasets: one naturally trained model, one model trained with IAAT, one model trained with MMA, and four adversarially trained models. All models are trained using Binary Cross Entropy with the default model architecture from CARLA, with three hidden layers of [18, 9, 3] units. The Adversarially Trained models were all trained with PGD at a variety of $\epsilon \in [0.05, 0.1, 0.15, 0.2]$. The IAAT model did not consider any hyperparameter choices, and the MMA model was trained using the original work's package [8] with the default hyperparameter choices.

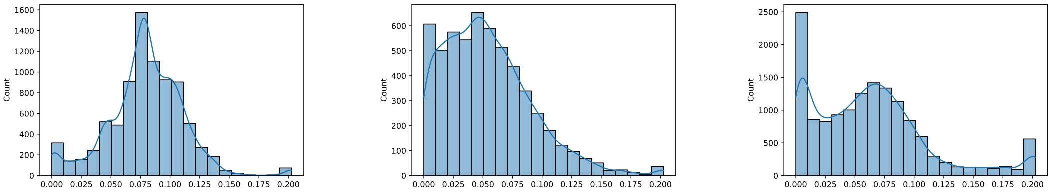
Recourse Methods I constructed Counterfactual Explanations for all models on a sample of 1000 negatively-classified test data points using three methods: Growing Spheres (GS), C-CHVAE, and SCFE. All hyperparameter choices for these methods were left as their CARLA defaults.

5.2 Metrics

To study the effects of the different training methods on accuracy, robustness, and recourse, I calculate the following metrics:

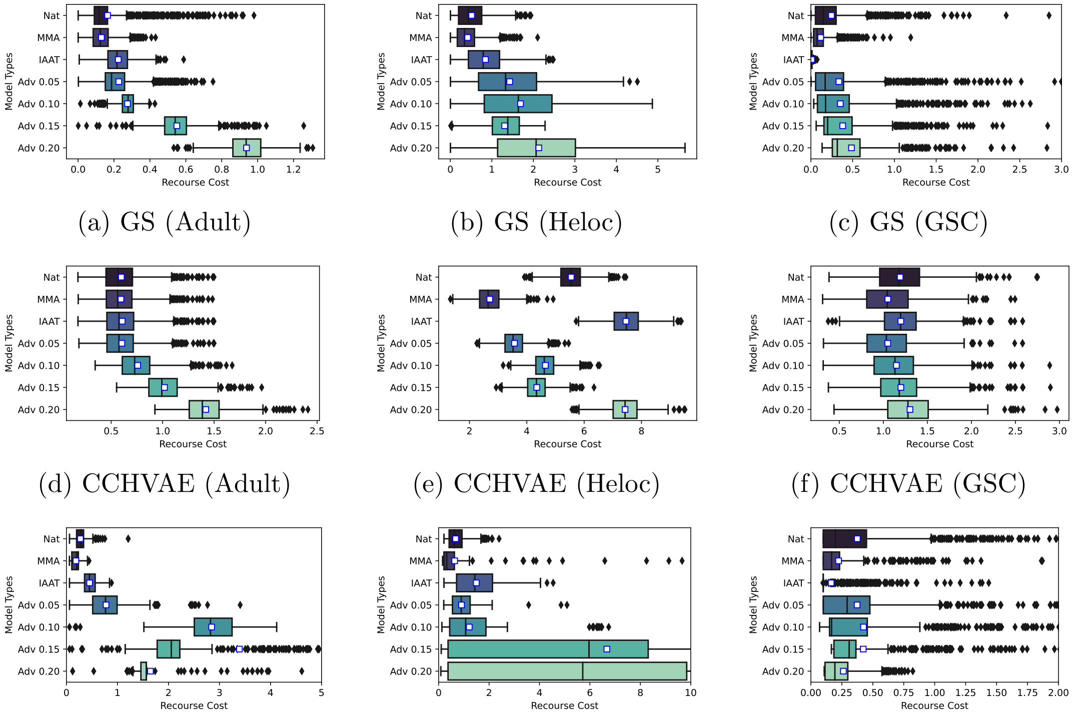
Standard Classification Performance A primary consideration in adversarial training is the trade-off in classification accuracy when compared with natural training. I record the standard classification accuracy of all models to measure the drop in accuracy that may accompany the different adversarial training methods. Formally, I measure: $\frac{1}{|\mathcal{D}_{test}|} \sum_{x_i \in \mathcal{D}_{test}} \mathbf{1}(f(x_i) = y_i)$. Given that I am experimenting with datasets with skewed target distributions, I also record the F1 score of each model on the minority target population.

Adversarial Success Rate Given that I am primarily concerned with the trade-off between robustness and recourse, and following the concept of "boundary error" introduced in [27] to disentangle standard performance and adversarial vulnerability, I also measure the success rate of adversarial attacks at various radii on the models. Formally, given an attack $\mathcal{A}_\epsilon$



(a) Adult (b) Heloc (c) GSC

Figure 5: AAT "Discovered" Radii Resulting from Adpative Adversarial Training

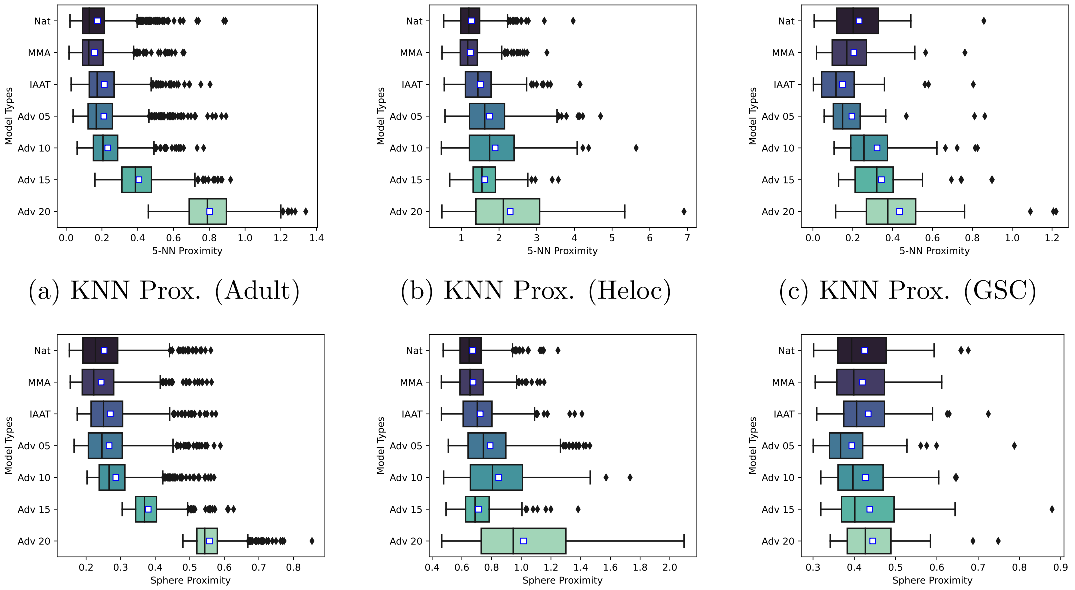


(a) GS (Adult) (b) GS (Heloc) (c) GS (GSC)

(d) CCHVAE (Adult) (e) CCHVAE (Heloc) (f) CCHVAE (GSC)

(g) SCFE (Adult) (h) SCFE (Heloc) (i) SCFE (GSC)

Figure 6: Recourse costs (defined as the ℓ_2 distance between a factual and counterfactual data point) for all methods and datsets. I observe that adaptive adversarial training shows significantly more competitive recourse costs than traditional adversarial training, and MMA training in particular shows almost no increase over natural training despite its robustness benefits.



(a) KNN Prox. (Adult) (b) KNN Prox. (Heloc) (c) KNN Prox. (GSC)

(d) Sphere Prox. (Adult) (e) Sphere Prox. (Heloc) (f) Sphere Prox. (GSC)

Figure 7: KNN and Sphere Manifold Proximity for Growing Spheres. I find that not only does adaptive adversarial training produce less expensive recourse than traditional adversarial training, but also recourse that is more faithful to the desired class these counterfactuals approximate.

such that $\mathcal{A}_\epsilon(x)$ identifies the most adversarial example on x within a radius ϵ, I measure $\frac{1}{|\mathcal{D}_{test}|}\sum_{x_i \in \mathcal{D}_{test}} \mathbf{1}(f(\mathcal{A}_\epsilon(x)) \neq f(x_i))$. I observe the adversarial success rate across the radii on which I train my traditional adversarial models. Note that this is an imperfect metric for measuring the success of AAT, as AAT assumes that some "attacks" at given radii represent real movements toward different classes; however, it is still useful to capture this information in considering the trade-off between traditional adversarial training and AAT.

Counterfactual Proximity The primary metric regarding recourse I am interested in observing is the ultimate recourse cost between the resultant models. As each specific domain's cost function is not concretely defined, I follow the convention of opting for ℓ_2 distance as a standard approximation. Formally, for each model I calculate: $\frac{1}{|\mathcal{D}_{test}|}\sum_{x_i \in \mathcal{D}_{test}} ||x_i^* - x_i||_2$, where x^* is the recourse calculated for x_i.

Manifold Proximity Motivated by the question of how faithful the resulting counterfactuals are to true movements towards the desired class, I estimate the distance between the counterfactuals each model produces and the desired class manifold these counterfactuals approximate.

I use two methods for this: a KNN distance measure and a sphere distance measure For KNN, I record the average ℓ_2 distance between the resulting counterfactuals and the five nearest neighbors of the desired class. For the sphere measure, I record the average ℓ_2 distance between the resulting counterfactuals and all neighbors of the desired class within an ℓ_2 ball of size ϵ, where ϵ is calculated as 20% of the average ℓ_2 distance between any two points in the dataset.

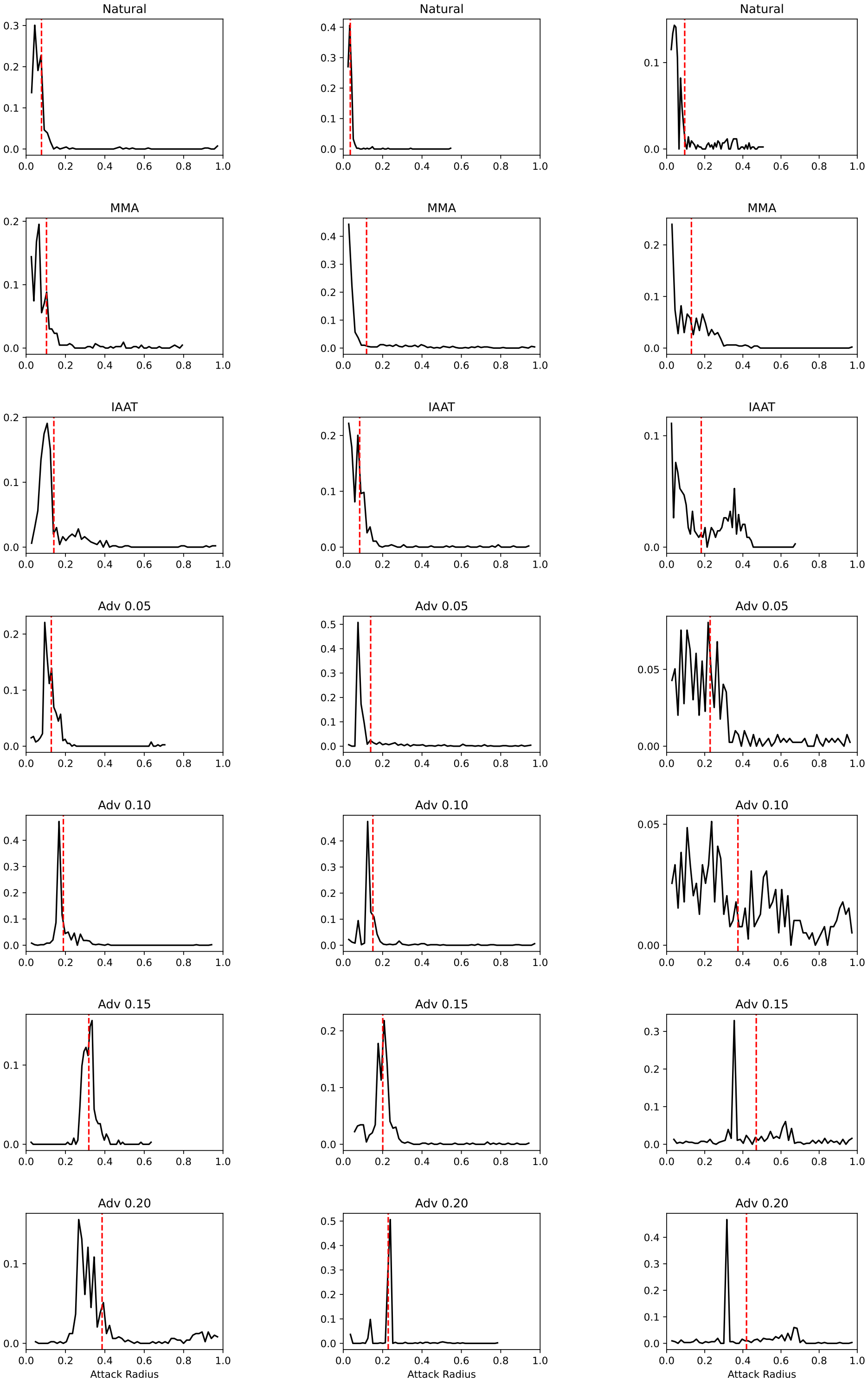
6 Results & Discussion

Standard Performance Figure 2 displays the classification accuracy and F1 scores of the various models. We can observe that for the Adult and Heloc datasets, adversarial training tends to decrease standard performance, with higher training radii correlating with worse performance. We can also observe that MMA training tends to keep performance consistent, and that IAAT worsens performance to a degree similar to adversarial training with an ϵ value between 0.05 and 0.1.

Robustness Figure 3 shows the vulnerability of the models under PGD attack at a variety of raddii ($\epsilon \in [0.05, 0.1, 0.15, 0.2, 0.25]$). We can observe that while traditional adversarial training creates substantially more robust models within a defined radius of attack, the degredation in robustness tends to be more severe among traditionally trained models than AAT methods when the radius increases beyond their predefined training threshold. MMA in particular shows surprisingly consistent robustness benfits, although they are more moderate than their adversarially trained counterparts'.

Counterfactual Proximity Figure 6 displays the cost of recourse across all datasets for the three recourse methods studied. We can observe consistently that adaptive adversarial training yields recourse with lower costs than traditional adversarial training, and in the case of MMA costs that are consistently competitive with natural training. This result seems unintuitive given the robustness benefits that MMA provides, and I believe this presents an interesting avenue for further research.

KNN & Sphere Manifold Distance Figure 7 shows the Manifold Proximity estimates for Growing Spheres across all datasets. Observe that adaptive adversarial training produces



(a) Adult

(b) GSC

(c) Heloc

Figure 8: Decision boundary proximity, estimated by the minimum successful PGD attack radius on a sample of 1000 instances. The height represents a proportion of the data, the average distance is shown in red.

recourse that is consistently closer to the desired class manifold than traditional adversarial training. This result, paired with the reduction in recourse costs, may suggest that adaptive adversarial training encourages more natural decision boundaries than traditional adversarial training, allowing for more meaningful recourse at lower costs.

Prevalence of Low Cost Recourse For recourse methods that optimize costs directly in the input space, I record the percentage of counterfactuals that have an ℓ_∞ cost less than 0.05 to measure the proportion of low cost recourse among the models. The results are recorded in Figure 4. Observe that adaptive adversarial training shows higher proportions of low cost recourse than traditional adversarially trained models; surprisingly, MMA training in particular finds proportions of low-cost recourse that are consistently competitive with natural training, despite its benefits in overall robustness.

Discovered Radii & Decision Boundary Distances Figure 5 displays the instance-wise discovered radii after AAT for all three datasets. We can observe that for all datasets, a variety of radii are found with unique distributions. This alludes to the fact that different underlying data distributions have different levels of inherent adversarial vulnerability, underscoring the challenge of estimating a proper singular radius at which to adversarially train. Figure 8 shows an estimation of the distribution of decision boundary proximities across all models, calculated by finding the minimum successful radius for PDG attack across a sample of 1000 instances. We can observe that traditional ϵ-adversarial training often limits proximity to the decision boundary $d > \epsilon_i$, while adaptive adversarial training shows a greater distribution in ultimate decision boundary proximties. In the case of MMA in particular, I find that the decision boundary proximities closely match that of the natural model, despite its improved robustness.

7 Conclusion

This work explores the effects of adaptive adversarial training on robustness and recourse, finding that it shows promising trade-offs between the two. I motivate my work with a observation of the effect of traditional adversarial training on recourse costs, and introduce scenarios under which adaptive adversarial training provides more affordable recourse. I conduct experiments on three datasets demonstrating that adaptive adversarial training yields significant robustness benefits over natural training with little cost incurred on recourse and standard performance, and provide evidence that adaptive adversarial training produces recourse that more faithfully represents movements towards the desired class manifold. Finally I analyze the resulting models' decision boundary margins, providing evidence that supports my observations on recourse costs under traditional adversarial training. I believe that adaptive adversarial training, and Max-Margin adversarial training in particular, presents a promising means of achieving the ultimate goals of robustness while preserving affordable recourse costs for end users.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness, 2021.

[2] Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets, 2019.

[3] Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. Imperceptible adversarial attacks on tabular data, 2019.

[4] Kieran Browne and Ben Swift. Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks. *arXiv preprint arXiv:2012.10076*, 2020.

[5] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018.

[6] Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness, 2020.

[7] Raphael Mazzine Barbosa de Oliveira and David Martens. A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences*, 11(16), 2021.

[8] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training, 2018.

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

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

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

[12] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), dec 2022.

[13] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. *CoRR*, abs/2002.06278, 2020.

[14] Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of adversarially robust models on algorithmic recourse, 2023.

[15] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, X. Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *ArXiv*, abs/1712.08443, 2017.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

[17] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis, 2021.

[18] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms, 2021.

[19] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. *Proceedings of The Web Conference 2020*, 2020.

[20] Mirka Saarela and Susanne Jauhiainen. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3, 02 2021.

[21] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[23] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019.

[24] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 284–293, New York, NY, USA, 2020. Association for Computing Machinery.

[25] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, 2020.

[26] Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*, 2017.

[27] Hongyang R. Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.