ALGORITHMIC RECOURSE IN THE FACE OF NOISY HUMAN RESPONSES

Martin Pawelczyk,* **Johannes van–den–Heuvel**
University of Tübingen
{first.last}@uni-tuebingen.de

Teressa Datta
Harvard University
tdatta.g@harvard.edu

Gjergji Kasneci
University of Tübingen
first.last@uni-tuebingen.de

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

ABSTRACT

As machine learning models are increasingly being deployed in high-stakes applications, there has been growing interest in providing recourse to individuals adversely impacted by model predictions (e.g., an applicant whose loan has been denied). To this end, several post hoc techniques have been proposed in recent literature. These techniques generate recourses under the assumption that the affected individuals will implement the prescribed recourses *exactly*. However, recent studies suggest that individuals often implement recourses in a noisy and inconsistent manner – e.g., raising their salary by \$505 if the prescribed recourse suggested an increase of \$500. Motivated by this, we introduce and study the problem of recourse invalidation in the face of noisy human responses. We propose a novel framework, EXPECTing noisy responses (EXPECT), which addresses the aforementioned problem by explicitly minimizing the probability of recourse invalidation in the face of noisy responses. Experimental evaluation with multiple real world datasets demonstrates the efficacy of the proposed framework, and supports our theoretical findings.

1 INTRODUCTION

Over the past decade, machine learning (ML) models are increasingly being deployed to make a variety of consequential decisions in domains such as finance, healthcare, and policy. Consequently, there is growing emphasis on designing tools and techniques which can provide *recourse* to individuals who have been adversely impacted by the predictions of these models (Voigt & Von dem Bussche, 2017). For example, when an individual is denied loan by a credit scoring model employed by a bank, they should be informed about the reasons for this decision and what can be done to reverse it. When providing a recourse to an affected individual, it is critical to ensure that the corresponding decision making entity (e.g., bank) is able to honor that recourse and approve any re-application that fully implements the recommendations outlined in the prescribed recourse (Wachter et al., 2018).

Several approaches in recent literature tackled the problem of providing recourses by generating *local* (instance level) counterfactual explanations (Wachter et al., 2018; Ustun et al., 2019; Karimi et al., 2020a; Poyiadzi et al., 2020; Van Looveren & Klaise, 2019). For instance, Wachter et al. (2018) proposed a gradient based approach which finds the nearest counterfactual resulting in the desired prediction.Ustun et al. (2019) proposed an integer programming based approach to obtain *actionable* recourses for linear classifiers. More recently, Karimi et al. (2021; 2020b) shed light on the spuriousness of the recourses generated by counterfactual/contrastive explanation techniques (Wachter et al., 2018; Ustun et al., 2019), and advocated for considering causal structure of the underlying data when generating recourses (Barocas et al., 2020; Mahajan et al., 2019; Pawelczyk et al., 2020).

The aforementioned approaches generate recourses under the assumption that the affected individuals will implement the prescribed recourses *exactly*. However, this may not always be the case in practice

*See https://arxiv.org/pdf/2203.06768.pdf for the complete version of this work.

e.g., if recourse prescribed for an individual suggests that they increase their salary by \$500, they may reapply for a loan with a salary increment of \$505 or even \$499.95. This phenomenon of *noisy responses to prescribed recourses* is very common in the real world, and has also been noted by Björkegren et al. (2020) who conducted a field experiment in Kenya by mimicking the "digital loan" setting to study algorithmic recourse in real-world scenarios. However, it is unclear if and how often such noisy implementations of recourses would result in positive outcomes for end users. This is due to the fact that there is no prior work which attempts to understand if recourses generated by state-of-the-art approaches will remain valid (i.e., result in positive outcomes) if they are implemented in a noisy manner (i.e., small changes are made to them). In fact, our analysis (Figure 3 in Appendix C) indicates that state-of-the-art approaches output recourses that are highly likely to be invalidated (as high as 60% of the time across multiple datasets) if small changes are made to the prescribed recourses. This poses a severe challenge to making algorithmic recourse practicable in the real world.

In this work, we introduce and address the critical problem of recourse invalidation in the face of noisy human responses. More specifically, we study if and how often recourses generated by state-of-the-art approaches become invalid (i.e., result in negative outcomes) if small changes are made to them, and provide a solution to address this problem.

2 Preliminaries

Notation Let $h : \mathcal{X} \to \mathcal{Y}$ denote a classifier which maps features $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ to labels $\mathcal{Y}$. Let $\mathcal{Y} = \{0, 1\}$ where 0 and 1 denote an unfavorable outcome (e.g., loan denied) and a favorable outcome (e.g., loan approved), respectively.

Since counterfactuals that propose changes to features such as gender are not actionable, we restrict the search space to ensure that only actionable changes are allowed. Let $\mathcal{A}$ denote the set of actionable counterfactuals. For a given predictive model h , and a predefined cost function $d_c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, the problem of finding a counterfactual explanation $\check{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ for an instance $\mathbf{x} \in \mathbb{R}^d$ can be expressed by the following optimization problem:

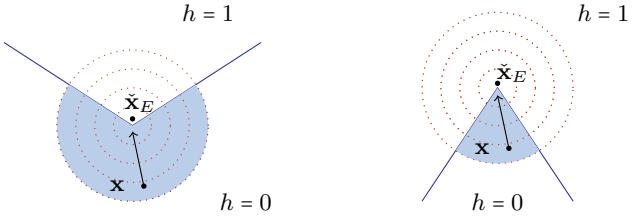
$$\check{\mathbf{x}} = \underset{\mathbf{x}' \in \mathcal{A}}{\arg\min}\ \mathcal{L}_1\big(h(\mathbf{x}'), 1\big) + \lambda \cdot d_c(\mathbf{x}, \mathbf{x}') \tag{1}$$

where $\lambda \geq 0$ is a trade-off parameter, and $\mathcal{L}_1(\cdot, \cdot)$ is the mean-squared-error (MSE) loss. The first term on the right-hand-side ensures that the model prediction corresponding to the counterfactual i.e., $h(\mathbf{x}')$ is close to the favorable outcome label 1. The second term encourages low-cost recourses; for example, Wachter et al. (2018) propose ℓ_1 or ℓ_2 distances to ensure that the distance between the original instance $\mathbf{x}$ and the counterfactual $\check{\mathbf{x}}$ is small.

2.1 Defining the Recourse Invalidation Rate

One of the key goals of this work is to understand if and when recourses output by state-of-the-art methods get invalidated when small changes are made to them. To this end, we formally define the notion of Recourse Invalidation Rate (IR) in this section.

We first introduce two key terms, namely, *prescribed recourses* and *implemented recourses*. A prescribed recourse is a recourse that was provided to an end user by some recourse method (e.g., increase salary by \$500). An implemented recourse corresponds to the recourse that the end user finally implemented (e.g., salary increment of \$505) upon being provided with the prescribed re-



(a) High IR (b) Low IR

Figure 1: Recourse invalidation rates (IRs) (blue area). The blue line indicates the decision boundary, $\mathbf{x}$ is the original input, and $\check{\mathbf{x}}_E$ is the generated recourse. The shaded region $(h = 0)$ belongs to the negative class, and the unshaded region $(h = 1)$ to the positive class. Figure 1a demonstrates the case where recourse invalidation rate for $\check{\mathbf{x}}_E$ is high because small perturbations to $\check{\mathbf{x}}_E$ are more likely to end up in the shaded region (negative class). Figure 1b demonstrates the case where recourse invalidation rate for $\check{\mathbf{x}}_E$ is low because small perturbations to $\check{\mathbf{x}}_E$ are less likely to end up in the shaded region.

course. With this basic terminology in place, we now proceed to formally define Recourse Invalidation Rate (IR) below.

Definition 1 (Recourse Invalidation Rate)**.** *For a given classifier h, the recourse invalidation rate corresponding to the counterfactual $\check{\mathbf{x}}_E = \mathbf{x} + \boldsymbol{\delta}_E$ output by a recourse method E is given by:*

$$\Delta(\check{\mathbf{x}}_E; \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\underbrace{h(\check{\mathbf{x}}_E)}_{\text{CF class}} - \underbrace{h(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon})}_{\text{class after response}}\Big], \tag{2}$$

where the expectation is taken with respect to a Gaussian random variable $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ which captures the noise in human responses.

Since the implemented recourses do not typically match the prescribed recourses, we add noise $\boldsymbol{\varepsilon}$ to the prescribed recourse $\check{\mathbf{x}}_E$. Since we primarily compute recourses for individuals $\mathbf{x}$ such that $h(\mathbf{x}) = 0$, the label corresponding to the counterfactual is given by $h(\check{\mathbf{x}}_E) = 1$ and therefore $\Delta \in [0, 1]$. For example, the following cases help understand our Recourse Invalidation Rate metric better: When $\Delta = 0$, then the prescribed recourse and the recourse implemented by the user agree all the time; and when $\Delta = 1$ then the prescribed recourse and the recourse implemented by the user never agree. Figure 1 provides further intuition about our IR metric.

3 Our Framework: EXPECT

In this section, we theoretically analyze the recourse invalidation rates (IRs) of state-of-the-art recourse methods. More specifically: 1) we provide a closed-form expression for the IR corresponding to any instance, 2) using the above closed-form expression, we analyze one of the most popular recourse methods (Wachter et al., 2018) proving that additional cost has to be incurred to generate robust recourses in the face of noisy human responses (Appendix A), and 3) we derive a general upper bound on the IR which is applicable to any valid recourse provided by any method with the underlying classifier being a differentiable model (Appendix A).

3.1 A Closed-Form Expression for Recourse Invalidation Rate

Before we introduce our formal result, we define $g(f(\mathbf{x}))$, where $f : \mathcal{X} \to \mathbb{R}$ is a differentiable scoring function (e.g., logit scoring function) and $g : \mathbb{R} \to \mathcal{Y}$ an activation function that maps logit scores to binary labels. Throughout the remainder of this work we will use $g(u) = \mathbb{I}[u > \xi]$, where ξ is a decision threshold in logit space. W.l.o.g. we will set $\xi = 0$. Here, we use definition 1 and provide a closed-form expression for the IR.

Theorem 1 (Closed-Form Recourse Invalidation Rate)**.** *A first-order approximation $\tilde{\Delta}$ to the recourse invalidation rate Δ in (2) under a Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ capturing the noise in human responses is given by:*

$$\tilde{\Delta}(\check{\mathbf{x}}_E; \boldsymbol{\Sigma}) = 1 - \Phi\left(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\right), \tag{3}$$

where Φ is the CDF of the univariate standard normal distribution $\mathcal{N}(0, 1)$ and $f(\check{\mathbf{x}}_E)$ denotes the logit score at $\check{\mathbf{x}}_E$ which is the recourse output by a recourse method E, and $h(\check{\mathbf{x}}_E) \in \{0, 1\}$.

This result is intuitive. First, when $f(\check{\mathbf{x}}_E) = 0$, then $\Delta = 0.5$ since $\Phi(0) = \frac{1}{2}$. This means that the prescribed recourse and the recourse implemented by the user *agree* 50% *of the time*. Second, when $f(\check{\mathbf{x}}_E) \to +\infty$, then $\Delta \to 0$ since $\Phi \to 1$, which means that the prescribed recourse and the recourse implemented by the user *always agree*. Finally, we consider the impact of the variance $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. If σ^2 decreases, then the size of the neighborhood where the recourse has to be robust shrinks, and therefore our IR $\Delta \to 0$ as $\sigma^2 \to 0$ if $f(\check{\mathbf{x}}_E) \geq 0$. The expression in (3) is a key ingredient required for both the algorithm presented next and our analysis in Appendix A.

3.2 Formulating and Optimizing our Objective

Our Objective The main idea is to find a recourse suggestion $\check{\mathbf{x}}$ whose prediction at any point y within some set around $\check{\mathbf{x}}$ belongs to the positive class with probability r. Hence, our idea consists of

Algorithm 1 `EXPECT`

Input: $\mathbf{x}$ s.t. $f(\mathbf{x}) < 0$, f, $\mathbf{\Sigma}$, $\lambda > 0$, Learning rate: $\alpha > 0$
Initialize: $\mathbf{x}' = \mathbf{x}$; $\tilde{\Delta} = \texttt{ClosedFormIR}(f, \mathbf{\Sigma}, \mathbf{x})$
while $\tilde{\Delta} > r$ **and** $f(\mathbf{x}') < 0$ **do**
  $\tilde{\Delta} = \texttt{ClosedFormIR}(f, \mathbf{\Sigma}, \mathbf{x}')$ {from Thm. 1}
  $\mathbf{x}' = \mathbf{x}' - \alpha \cdot \nabla_{\mathbf{x}'} \mathcal{L}(\mathbf{x}'; \mathbf{\Sigma}, r, \lambda)$ {Optimize (5)}
end while
Return: $\check{\mathbf{x}} = \mathbf{x}'$

minimizing the recourse invalidation rate subject to the constraint of low cost recourse. Our objective looks as follows:

$$\check{\mathbf{x}} = \underset{\mathbf{x}' \in \mathcal{A}}{\arg\min}\, \Delta(\mathbf{x}'; \mathbf{\Sigma}) \text{ s.t. } d_c(\mathbf{x}, \mathbf{x}') \leq q \wedge h(\mathbf{x}') \neq 0, \tag{4}$$

where q is a cost budget, $\Delta(\mathbf{x}'; \mathbf{\Sigma})$ is the recourse invalidation rate from (1), d_c measures the distance between the factual input and the prescribed recourse, and h is the fixed classifier. We use a Lagrangian formulation with parameter λ to encourage balance between the different objectives as follows:

$$\mathcal{L}(\mathbf{x}'; \mathbf{\Sigma}, r, \lambda) = \mathcal{L}_0(\mathbf{x}'; \mathbf{\Sigma}, r) + \mathcal{L}_1\big(f(\mathbf{x}'), s)\big) + \lambda \cdot d_c(\mathbf{x}', \mathbf{x}), \tag{5}$$

where $\mathcal{L}_0 = \max(0, \Delta(\mathbf{x}'; \mathbf{\Sigma}) - r)$ and r is the target IR. The new component $\mathcal{L}_0$ is a Hinge loss encouraging that the prescribed recourse has a low probability of invalidation, and the parameter $\mathbf{\Sigma}$ controls the shape and the size of the neighbourhood in which the recourse has to be robust in line with Definition 1. In practical use-cases the choice of r would depend on the risk-aversion of the end-user. If the end-user is not confident about achieving a 'precision landing', then a rather low invalidation target should be chosen (i.e., $r < 0.5$). In the extreme case, when $r = 0$, the objective would encourage finding recourses that always lead to a positive outcome for a given neighborhood shape and size controlled by $\mathbf{\Sigma}$.

Optimization We suggest two ways to minimize the objective in (5). First, we can approximate the IR in (2) (i.e., Δ) by replacing it with the approximate closed-form IR expression $\tilde{\Delta}$ from (3) and minimize (5). Algorithm 1 then proceeds in an iterative fashion where we do gradient descent on the loss function in (5). This procedure is executed repeatedly until the class label flips from 0 to 1 and $\tilde{\Delta}$ is less than or equal to r. Second, instead of using (3) we can use a Monte-Carlo approach in combination with the *reparametrization trick* to obtain a differentiable approximation of IR. Intuitively, this trick separates the randomness of the noise distribution and the influence of the distribution parameters with respect to which we want to take the gradients. We refer to Kingma & Welling (2013) for a detailed discussion of this trick.

Synthetic Example In Figure 2, we demonstrate how `EXPECT` finds recourses relative to `Wachter`'s algorithm. We see that `EXPECT` finds recourses in line with both the chosen invalidation target (e.g., in the left panel the target is set $r = 0.3$) and the variance σ^2 which controls the size of the neighborhood, in which the recourses have to be robust.

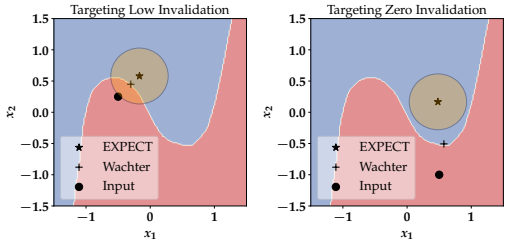


Figure 2: Computing recourses with low IRs on the binary classification *moon data set* (Pedregosa et al., 2011) for a DNN classifier with 100 hidden units. The circles around `EXPECT`'s recourses have radius 2σ, i.e., they show the region where 95% of recourse inaccuracies fall when $\sigma^2 = 0.05$. **Left**: We chose an invalidation target of $r = 0.3$, i.e., 30% of the recourse responses would fail under spherical response inaccuracies $\varepsilon \sim \mathcal{N}(\mathbf{0}, 0.05 \cdot \mathbf{I})$. **Right**: The same setup, but now we chose $r = 0$.

4 Empirical Results

We now present our empirical analysis. We study the effectiveness of `EXPECT` at finding robust recourses in the presence of noisy human responses.

4.1 Experimental Setup

We first describe the synthetic and real-world data sets used. We then describe the predictive models employed in our experiments, and the state-of-the-art recourse methods that we use as baselines.

Real-World Data and Noisy Responses Regarding real-world data, we use the same data sets as provided in the recourse and counterfactual explanation library `CARLA` (Pawelczyk et al., 2021). The *Adult* data set Dua & Graff (2017) originates from the 1994 Census database, consisting of 14 attributes and 48,842 instances. The class label indicates whether an individual has an income greater than 50,000 USD/year. The *Give Me Some Credit* (GMC) data set Kaggle-Competition (2011) is a credit scoring data set, consisting of 150,000 observations and 11 features. The class label indicates if the corresponding individual will experience financial distress within the next two years (*SeriousDlqin2yrs* is 1) or not. The *COMPAS* data set Angwin et al. (2016) contains data for more than 10,000 criminal defendants in Florida. It is used by the jurisdiction to score defendant's likelihood of re-offending. The class label indicates if the corresponding defendant is high or low risk for recidivism. All the data sets were normalized so that $\mathbf{x} \in [0,1]^d$. Across all experiments, we add noise $\boldsymbol{\varepsilon}$ to the prescribed recourse $\check{\mathbf{x}}_E$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$ and $\sigma^2 \in \{0.01, 0.025, 0.05\}$.

Methods We compare the recourses generated by `EXPECT` to three different methods which aim to generate low-cost recourses using fundamentally different principles: `AR (-LIME)` uses an integer-programming-based objective (Ustun et al., 2019), `Wachter` uses a gradient-based objective (Wachter et al., 2018), and `GS` is based on a random search algorithm (Laugel et al., 2017). We have used the recourse method implementations from the `CARLA` library (Pawelczyk et al., 2021).

Prediction Models For all data sets (except the synthetic one), we trained both ReLU-based NN models with 50 hidden layers and logistic regression classifiers.

		Adult				Compas				GMC			
		AR	Wachter	GS	EXPECT	AR	Wachter	GS	EXPECT	AR	Wachter	GS	EXPECT
LR	RA (↑)	0.98	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	AIR (↓)	0.5 ± 0.01	0.46 ± 0.02	0.35 ± 0.11	$\mathbf{0.22 \pm 0.02}$	0.48 ± 0.04	0.47 ± 0.02	0.3 ± 0.18	$\mathbf{0.08 \pm 0.01}$	0.47 ± 0.06	0.45 ± 0.03	0.48 ± 0.04	$\mathbf{0.24 \pm 0.01}$
	AC (↓)	$\mathbf{0.55 \pm 0.4}$	0.62 ± 0.43	2.12 ± 1.05	1.24 ± 0.81	$\mathbf{0.16 \pm 0.17}$	0.22 ± 0.17	0.73 ± 0.45	0.66 ± 0.27	$\mathbf{0.29 \pm 0.27}$	0.49 ± 0.51	$\mathbf{0.28 \pm 0.31}$	1.01 ± 2.01
NN	RA (↑)	0.38	**1.0**	**1.0**	**1.0**	0.84	**1.0**	**1.0**	**1.0**	0.38	**1.0**	**1.0**	**1.0**
	AIR (↓)	0.49 ± 0.03	0.5 ± 0.02	0.48 ± 0.02	$\mathbf{0.28 \pm 0.01}$	0.34 ± 0.09	0.46 ± 0.02	0.43 ± 0.07	$\mathbf{0.18 \pm 0.03}$	0.34 ± 0.07	0.43 ± 0.03	0.45 ± 0.03	$\mathbf{0.27 \pm 0.03}$
	AC (↓)	1.05 ± 0.22	$\mathbf{0.3 \pm 0.19}$	2.99 ± 1.51	1.43 ± 0.49	1.15 ± 0.52	$\mathbf{0.2 \pm 0.16}$	0.81 ± 0.45	0.8 ± 0.34	0.2 ± 0.19	0.26 ± 0.18	$\mathbf{0.12 \pm 0.09}$	0.47 ± 0.21

Table 1: Recourse accuracy (RA), average recourse invalidation rate (AIR) for $\sigma^2 = 0.01$ and average cost (AC) across different recourse methods. Recourses that use our framework `EXPECT` are more robust compared to those produced by existing baselines. For `EXPECT`, we generated recourses by setting $r = 0.35$ and $\sigma^2 = 0.01$. Therefore, the AIR should be at most 0.35; in line with our results.

4.2 Evaluating the EXPECT Framework

Measures We consider three measures in our evaluation: 1) We measure the *average cost* (AC) required to act upon the prescribed recourses where the average is taken with respect to the instances in the test set for which a given method provides recourse. Since the algorithms are optimizing for the ℓ_1-norm we use this as our cost measure. 2) We use *recourse accuracy* (RA) defined as the fraction of instances in the test set for which acting upon the prescribed recourse results in the desired prediction. 3) We compute the *average IR* across every instance in the test set. To do that, we sample 10,000 points from $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ for every instance and compute IR in (2). Then the *average IR* quantifies recourse robustness where the individual IRs are averaged over all instances from the test set.

Results Here, we evaluate the robustness, costs and recourse accuracy of the recourses generated by our framework `EXPECT` relative to the baselines. We consider a recourse robust if the recourse remains valid (i.e., results in positive outcome) even after small changes are made to it (i.e., humans implement it in a noisy manner). Table 1 shows the average IR for different methods across different real world data sets and classifiers when $\sigma^2 = 0.01$. It can be seen that `EXPECT` has the lowest invalidation rate across all real-world data sets and classifiers. We also consider if the robustness achieved by our framework is associated with an additional cost i.e., by sacrificing recourse accuracy (RA) or by increasing the average recourse cost (AC). We compute AC of the recourses output by all the algorithms on various data sets and find that `EXPECT` usually has the highest or second highest recourse costs, while the recourse accuracy is at 100% across classifiers and data sets.

5 CONCLUSION

In this work, we introduced and studied the critical problem of recourse invalidation in the face of noisy human responses. We proposed a novel framework, EXPECTing noisy responses (EXPECT), which addresses the aforementioned problem by explicitly minimizing the probability of recourse invalidation in the face of noisy responses.

REFERENCES

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, 2016.

Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, New York, NY, USA, 2020. ACM.

Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-proof machine learning. *arXiv:2004.03865*, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Kaggle-Competition. "give me some credit data set", 2011.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020a.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020 (WWW)*. ACM, 2020.

Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Advances in Neural Information Processing Systems (NeurIPS) (Benchmark and Datasets Track)*, volume 34, 2021.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens ofadversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 2011.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 344–350, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375850. URL https://doi.org/10.1145/3375627.3375850.

Berk Ustun, Alexander Spangher, and Y. Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

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

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2018.

A Theoretical Analysis

A.1 Recourse Invalidation Rate for Wachter et al. (2018)

Next, we specify the recourse invalidation rate for the algorithm proposed by Wachter et al. (2018). For their algorithm, Pawelczyk et al. (2022) give a closed-form recourse solution for logistic regression classifiers when $d_c = \|\mathbf{x} - \mathbf{x}'\|_2$ and the MSE-loss is used. Then the solution takes the following form: $\check{\mathbf{x}}_{\text{Wachter}}(s) = \mathbf{x} + \frac{s - f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2^2} \nabla f(\mathbf{x})$, where s is the target logit score. More specifically, to arrive at the desired class with probability of 0.5, the target score for a sigmoid function is $s = 0$, where the logit corresponds to a 0.5 probability for $y = 1$. The next statement quantifies the IR of recourses output by Wachter et al. (2018).

Lemma 1. *For the logistic regression classifier, consider the recourse output by Wachter et al. (2018):* $\check{\mathbf{x}}_{\text{Wachter}}(s) = \mathbf{x} + \frac{s - f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2^2} \nabla f(\mathbf{x})$. *Then the recourse invalidation rate has the following closed-form:*

$$\Delta(\check{\mathbf{x}}_{\text{Wachter}}(s); \sigma^2 \mathbf{I}) = 1 - \Phi\left(\frac{s}{\sigma \|\nabla f(\mathbf{x})\|_2} \right), \tag{6}$$

where s is the target logit score.

Proof Sketch. The proof uses the recourse expression from Pawelczyk et al. (2022) and plugs it into (3). □

A recourse generated by Wachter et al. (2018) such that $f(\check{\mathbf{x}}_{\text{Wachter}}) = s = 0$ will result in $\Delta = 0.5$. Note that this is true regardless of the choice of $\mathbf{\Sigma}$. To obtain recourse that is more robust to noisy responses from users, i.e., $\Delta \to 0$, the decision maker can choose a higher logit target score of $s' > s \geq 0$ since this decreases the recourse invalidation rate, i.e., $\Delta(\check{\mathbf{x}}_{\text{Wachter}}(s)) > \Delta(\check{\mathbf{x}}_{\text{Wachter}}(s'))$.[1] Thus, we can now see that there exists a trade-off between robustness to noisy human responses and cost: since $\|\check{\mathbf{x}}_{\text{Wachter}}(s')\| > \|\check{\mathbf{x}}_{\text{Wachter}}(s)\|$ while $\Delta(\check{\mathbf{x}}_{\text{Wachter}}(s')) < \Delta(\check{\mathbf{x}}_{\text{Wachter}}(s))$, we see that a higher target score leads to a more robust recourse, while increasing the recourse costs holding all other variables constant (e.g. $\mathbf{x}$ and σ^2).

A.2 A General Upper Bound on the Recourse Invalidation Rate

Next, we derive a general upper bound on the recourse invalidation rate. This bound is applicable to any method E that provides recourses resulting in a positive outcome.

[1] This is not generally true in non-linear models.

Lemma 2. *Let $\check{\mathbf{x}}_E$ be the output produced by some recourse method E such that $h(\check{\mathbf{x}}_E) = 1$. Then, an upper bound on $\tilde{\Delta}$ from (3) is given by:*

$$\tilde{\Delta}(\check{\mathbf{x}}_E; \sigma^2\mathbf{I}) \leq 1 - \Phi\left(c + \frac{\omega}{\sigma}\frac{\|\nabla f(\mathbf{x})\|_2}{\|\nabla f(\check{\mathbf{x}}_E)\|_2}\frac{\|\boldsymbol{\delta}_E\|_1}{\sqrt{\|\boldsymbol{\delta}_E\|_0}}\right), \tag{7}$$

where $c = \frac{f(\mathbf{x})}{\sigma \cdot \|\nabla f(\mathbf{x})\|_2}$ is a constant, $\boldsymbol{\delta}_E = \check{\mathbf{x}}_E - \mathbf{x}$, and $\omega > 0$ is the cosine of the angle between the vectors $\nabla f(\mathbf{x})$ and $\boldsymbol{\delta}_E$.

Proof Sketch. Starting from (3), we use the approximation of $f(\check{\mathbf{x}}_E)$ in combination with the fact that $\omega = \frac{\nabla f(\mathbf{x})^T \boldsymbol{\delta}_E}{\|\boldsymbol{\delta}_E\|_2 \cdot \|\nabla f(\mathbf{x})\|_2}$. We conclude using a *lower bound* on the ℓ_2-norm of $\boldsymbol{\delta}_E$ which depends on both its ℓ_0 and ℓ_1-norms. □

The right term in the inequality entails that the upper bound depends on the ratio of the ℓ_1 and ℓ_0-norms of the recourse action $\boldsymbol{\delta}_E$ provided by recourse method E. The higher the ℓ_1/ℓ_0 ratio of the recourse actions, the tighter the bound. The bound is tight when $\|\boldsymbol{\delta}_E\|_0$ assumes minimum value i.e., $\|\boldsymbol{\delta}_E\|_0 = 1$ since at least one feature needs to be changed to flip the model prediction.

B Proofs

B.1 Proof of Theorem 1

Theorem 1. *A first-order approximation $\tilde{\Delta}$ to the recourse invalidation rate Δ in (2) under a Gaussian distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ capturing the noise in human responses is given by:*

$$\tilde{\Delta}(\check{\mathbf{x}}_E; \boldsymbol{\Sigma}) = 1 - \Phi\left(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\right), \tag{8}$$

where Φ is the CDF of the univariate standard normal distribution $\mathcal{N}(0,1)$, $f(\check{\mathbf{x}}_E)$ denotes the logit score at $\check{\mathbf{x}}_E$ which is the recourse output by a recourse method E, and $h(\check{\mathbf{x}}_E) \in \{0,1\}$.

Proof. Let the random variable $\boldsymbol{\varepsilon}$ follow a multivariate normal distribution, i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The following result is a well-known fact: $\mathbf{v}^T\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{v}^T\boldsymbol{\mu}, \mathbf{v}\boldsymbol{\Sigma}\mathbf{v}^T)$ where $\mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{x}$ denote the input sample for which we wish to find a counterfactual $\check{\mathbf{x}}_E = \mathbf{x} + \boldsymbol{\delta}_E$. Recall from Definition 1 that we have to evaluate:

$$\begin{aligned}\Delta &= \mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\underbrace{h(\check{\mathbf{x}}_E)}_{\text{CE class}} - \underbrace{h(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon})}_{\text{class after response}}\Big] \\ &= 1 - \mathbb{E}_{\boldsymbol{\varepsilon}}\big[h(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon})\big],\end{aligned} \tag{9}$$

where we have used that the first term is a constant and evaluates to 1 by the definition of a counterfactual explanation. It remains to evaluate the expectation: $\mathbb{E}_{\boldsymbol{\varepsilon}}\big[h(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon})\big]$. Next, we note that (9) can equivalently be expressed in terms of the logit outcomes:

$$\Delta = \mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\underbrace{\mathbb{I}\big[f(\check{\mathbf{x}}_E) > 0\big]}_{\text{CE class}} - \underbrace{\mathbb{I}\big[f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) > 0\big]}_{\text{class after perturbation}}\Big] = \Big(1 - \mathbb{E}_{\boldsymbol{\varepsilon}}\big[\mathbb{I}\big[f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) > 0\big]\big]\Big). \tag{10}$$

Again, we are interested in the second term, which evaluates to:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\big[\mathbb{I}\big[f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) > 0\big]\big] = 0 \cdot \mathbb{P}\Big(f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) < 0\Big) + 1 \cdot \mathbb{P}\Big(f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) > 0\Big). \tag{11}$$

Next, consider the first-order Taylor approximation: $f(\check{\mathbf{x}}_E + \boldsymbol{\varepsilon}) \approx f(\check{\mathbf{x}}_E) + \nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}$. Hence, we know $\nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}$ approximately follows $\mathcal{N}(\mathbf{0}, \nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T)$. Now, the second term can be

computed as follows:

$$\mathbb{P}\Big(f(\check{\mathbf{x}}_E+\boldsymbol{\varepsilon})>0\Big) \approx \mathbb{P}\Big(f(\check{\mathbf{x}}_E)>-\nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}\Big) = \mathbb{P}\Big(-f(\check{\mathbf{x}}_E)<\nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}\Big) \tag{12}$$

$$= 1-\mathbb{P}\Big(-f(\check{\mathbf{x}}_E)>\nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}\Big) \tag{13}$$

$$= 1-\mathbb{P}\Bigg(\underbrace{\frac{\nabla f(\check{\mathbf{x}}_E)^T\boldsymbol{\varepsilon}}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}}_{\text{Mean 0 Gaussian RV}} < \underbrace{-\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}}_{\text{Constant}}\Bigg)$$

$$= 1-\Phi\Bigg(-\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\Bigg)$$

$$= \Phi\Bigg(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\Bigg), \tag{14}$$

where the last line follows due to symmetry of the standard normal distribution (i.e., $\Phi(-u) = 1-\Phi(u)$). Putting the pieces together, we have:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\mathbb{I}\Big[f(\check{\mathbf{x}}_E+\boldsymbol{\varepsilon})>0\Big]\Big] = 0\cdot\mathbb{P}\Big(f(\check{\mathbf{x}}_E+\boldsymbol{\varepsilon})<0\Big)+1\cdot\mathbb{P}\Big(f(\check{\mathbf{x}}_E+\boldsymbol{\varepsilon})\geq 0\Big) \tag{15}$$

$$= \Phi\Bigg(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\Bigg). \tag{16}$$

Thus, we have:

$$\Delta \approx \tilde{\Delta} = 1-\Phi\Bigg(\frac{f(\check{\mathbf{x}}_E)}{\sqrt{\nabla f(\check{\mathbf{x}}_E)\boldsymbol{\Sigma}\nabla f(\check{\mathbf{x}}_E)^T}}\Bigg), \tag{17}$$

which completes our proof. Note that this is equivalent to $\mathbb{P}\Big(f(\check{\mathbf{x}}_E+\boldsymbol{\varepsilon})<0\Big)$, and thus we are "counting" how often perturbations to $\check{\mathbf{x}}_E$ sampled from $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ result in flips back to the undesired class. □

C Empirical Results

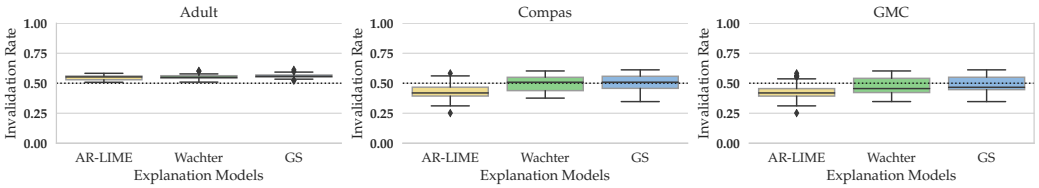


Figure 3: Boxplots of recourse invalidation rates (IR) across generated recourses $\check{\mathbf{x}}$ from the test set for Neural Network classifiers on three data sets. The recourses were generated by three different recourse methods (i.e., `AR-LIME`, `Wachter`, and `GS`) using `CARLA` (Pawelczyk et al., 2021). These methods use different techniques (i.e., integer programming, gradient search, and random search) to find minimum cost recourses. The noisy human responses were simulated by adding small Gaussian random noise $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},0.05\cdot\mathbf{I})$ to $\check{\mathbf{x}}$, where $\mathbf{I}$ is the identity matrix. The recourse invalidation rates are as high as 60% across all datasets.