

# Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation

Junqi Jiang<sup>1</sup>, Jianglin Lan<sup>2</sup>, Francesco Leofante<sup>1</sup>, Antonio Rago<sup>1</sup>, Francesca Toni<sup>1</sup>

<sup>1</sup> Department of Computing, Imperial College London

<sup>2</sup> James Watt School of Engineering, University of Glasgow

{junqi.jiang, f.leofante, a.rago, f.toni}@imperial.ac.uk, jianglin.lan@glasgow.ac.uk

## Abstract

Counterfactual Explanations (CEs) have received increasing interest as a major methodology for explaining neural network classifiers. Usually, CEs for an input-output pair are defined as data points with minimum distance to the input that are classified with a different label than the output. To tackle the established problem that CEs are easily invalidated when model parameters are updated (e.g. retrained), studies have proposed ways to certify the robustness of CEs under model parameter changes bounded by a norm ball. However, existing methods targeting this form of robustness are not sound or complete, and they may generate implausible CEs, i.e., outliers wrt the training dataset. In fact, no existing method simultaneously optimises for proximity and plausibility while preserving robustness guarantees. In this work, we propose Provably RObust and PLAusible Counterfactual Explanations (PROPLACE)<sup>1</sup>, a method leveraging on robust optimisation techniques to address the aforementioned limitations in the literature. We formulate an iterative algorithm to compute provably robust CEs and prove its convergence, soundness and completeness. Through a comparative experiment involving six baselines, five of which target robustness, we show that PROPLACE achieves state-of-the-art performances against metrics on three evaluation aspects.

## 1 Introduction

Counterfactual Explanations (CEs) have become a major methodology to explain NNs due to their simplicity, compliance with the regulations [Wachter *et al.*, 2017], and alignment with human thinking [Celar and Byrne, 2023]. Given an input point to a classifier, a CE is a modified input classified with another, often more desirable, label. Consider a customer that is denied a loan by the machine learning

system of a bank. A CE the bank provided for this customer could be, *the loan application would have been approved, had you raised your annual salary by \$ 6000*. Several desired properties of CEs have been identified in the literature, the most fundamental of which is *validity*, requiring that the CE needs to be correctly classified with a specified label [Tolomei *et al.*, 2017]. *Proximity* refers to the closeness between the CE and the input measured by some distance metric, which translates to a measure of the effort the end user has to make to achieve the prescribed changes [Wachter *et al.*, 2017]. The CEs should also lie on the data manifold of the training dataset and not be an outlier, which is assessed via *plausibility* [Poyiadzi *et al.*, 2020]. Most recently, the *robustness* of CEs, amounting to their validity under various types of uncertainty, has drawn increasing attention due to its real-world importance. In this work, we consider robustness to the model parameter changes occurring in the classifier on which the CE was generated. Continuing the loan example, assume the bank’s machine learning model is retrained with new data, while, in the meantime, the customer has achieved a raise in salary (as prescribed by the CE). The customer may then return to the bank only to find that the previously specified CE is now invalidated by the new model. In this case, the bank could be seen as being responsible by the user and could potentially be legally liable, risking financial and reputational damage to the organisation. The quality of such unreliable CE is also questionable: [Rawal *et al.*, 2020; Dutta *et al.*, 2022] have shown that CEs found by existing non-robust methods are prone to such invalidation due to their closeness to the decision boundary.

Various methods have been proposed to tackle this issue. [Nguyen *et al.*, 2022; Dutta *et al.*, 2022; Hamman *et al.*, 2023] focus on building heuristic methods using model confidence, Lipschitz continuity, and quantities related to the data distribution. [Upadhyay *et al.*, 2021; Black *et al.*, 2022; Jiang *et al.*, 2023] consider optimising the validity of CEs under bounded model parameter changes, which are also empirically shown to be robust to the unbounded parameter changes scenarios. Among the existing methods, only [Jiang *et al.*, 2023] provides robustness guarantees in a formal approach, which are known to be lacking in the explainable AI (XAI) literature in general, aside from some notable examples, e.g. as introduced in [Marques-Silva and Ignatiev, 2022]. Their method generates

<sup>1</sup>The implementation is available at <https://github.com/junqi-jiang/proplace>

such provably robust CEs via iteratively tuning the hyperparameters of an arbitrary non-robust CEs method and testing for robustness. However, this method cannot always guarantee soundness and is not complete, which is also the case for the method in [Upadhyay *et al.*, 2021]. Another limitation in the current literature is that the methods targeting this form of robustness guarantee do not find plausible CEs, limiting their practical applicability.

Such limitations have motivated this work. After discussing relevant studies in Section 2, we introduce the robust optimisation problem for computing CEs with proximity property as the objective, and robustness and plausibility properties as constraints (Section 3). In Section 4, we then present Provably ROBust and PLAusible CEs (PROPLACE), a method leveraging on robust optimisation techniques to address the limitation in the literature that no method optimises for proximity and plausibility while providing formal robustness guarantees. We show the (conditional) soundness and completeness of our method, and give a bi-level optimisation procedure that will converge and terminate. Finally, in our experiments, we compare PROPLACE with six existing CE methods, five of which target robustness, on four benchmark datasets. The results show that our method achieves the best robustness and plausibility, while demonstrating superior proximity among the most robust baselines.

## 2 Related Work

As increasing interest has been focused on XAI, a plethora of CE generation methods have been proposed (see [Guidotti, 2022] for a recent overview). Given our focus on neural networks, we cover those explaining the outputs of these models. [Wachter *et al.*, 2017] proposed a gradient-based optimisation method targeting the validity and proximity of CEs. Similarly, using the mixed integer programming (MILP) representation of neural networks, [Mohammadi *et al.*, 2021] formulated the CEs search into a constrained optimisation problem such that the resulting CEs are guaranteed to be valid. [Mothilal *et al.*, 2020] advocated generating a diverse set of CEs for each input to enrich the information provided to the explaine. Several works also addressed *actionability* constraints [Ustun *et al.*, 2019; Verma *et al.*, 2022; Vo *et al.*, 2023], only allowing changes in the actionable features of real users. [Poyiadzi *et al.*, 2020] proposed a graph-based method to find a path of CEs that are all lying within the data manifold. Several other works have proposed to use (variational) auto-encoders or nearest neighbours to induce plausibility [Dhurandhar *et al.*, 2018; Pawelczyk *et al.*, 2020a; Van Looveren and Klaise, 2021]. Among these properties, actionability and plausibility are two orthogonal considerations which make the CEs realistic in practice, and trade-offs have been identified between plausibility and proximity [Pawelczyk *et al.*, 2020b].

In this work, our focus is on the property of robustness to changes in the model parameters, i.e. weights and biases in the underlying classifier. Several studies looked at CEs under bounded model parameter changes of a neural network: [Upadhyay *et al.*, 2021] formulated a

novel loss function and solved using gradient-based methods. [Black *et al.*, 2022] proposed a heuristic based on the classifier’s Lipschitz constant and the model confidence to search for robust CEs. [Jiang *et al.*, 2023] used interval abstractions [Prabhakar and Afzal, 2019] to certify the robustness against bounded parameter changes, and embed the certification process into existing CE methods. Differently to our approach, these methods do not generate plausible CEs or guarantee that provably robust CEs are found. Other relevant works place their focus on the robustness of CEs against unbounded model changes. [Ferrario and Loi, 2022] took the approach of augmenting the training data with previously generated CEs. [Nguyen *et al.*, 2022] focused on the data distribution and formulated the problem as posterior probability ratio minimisation to generate robust and plausible CEs. By using first- and second-moment information, [Bui *et al.*, 2022] proposed lower and upper bounds on the CEs’ validity under random parameter updates and generated robust CEs using gradient descent. [Dutta *et al.*, 2022] defined a novel robustness measure based on the model confidences over the neighbourhood of the CE, and used dataset points that satisfy some robustness test to find close and plausible CEs. Their notion is then further re-calibrated for neural networks with probabilistic robustness guarantees in [Hamman *et al.*, 2023]. Trade-offs between robustness and proximity were discussed by [Pawelczyk *et al.*, 2022] and [Upadhyay *et al.*, 2021].

Other forms of CEs’ robustness have also been investigated, for example, robustness against: input perturbations [Alvarez-Melis and Jaakkola, 2018; Sharma *et al.*, 2020; Bajaj *et al.*, 2021; Dominguez-Olmedo *et al.*, 2022; Huai *et al.*, 2022; Virgolin and Fracaros, 2023; Zhang *et al.*, 2023]; noise in the execution of CEs [Pawelczyk *et al.*, 2022; Leofante and Lomuscio, 2023b; Leofante and Lomuscio, 2023a; Maragno *et al.*, 2023]; and model multiplicity [Pawelczyk *et al.*, 2020b; Leofante *et al.*, 2023].

## 3 Preliminaries and Problem Statement

**Notation.** Given an integer  $k$ , we use  $[k]$  to denote the set  $\{1, \dots, k\}$ . We use  $|S|$  to denote the cardinality of a set  $S$ .

**Neural Network (NN).** We denote a NN as  $\mathcal{M}_\Theta : \mathcal{X} \subseteq \mathbb{R}^d \rightarrow \mathcal{Y} \subseteq \mathbb{N}$ , where the inputs are  $d$ -dimensional vectors and the outputs are discrete class labels.  $\Theta$  represents the collection of parameters that characterise the NN. Throughout the paper, we will illustrate our method using the binary classification case (i.e.  $\mathcal{Y} = \{0, 1\}$ ), though the method is readily applicable to multi-class classification. Let  $\mathcal{M}_\Theta(x)$  also (with an abuse of notation) refer to the pre-sigmoid (logit) value in the NN. Then, for an input  $x \in \mathcal{X}$ , we say  $\mathcal{M}_\Theta$  classifies  $x$  as class 1 if  $\mathcal{M}_\Theta(x) \geq 0$ , otherwise  $\mathcal{M}_\Theta$  classifies  $x$  as class 0.

**Counterfactual Explanation (CE).** For an input  $x \in \mathcal{X}$  that is classified to the unwanted class 0 (assumed throughout the paper), a CE  $x' \in \mathcal{X}$  is some other data point “similar” to the input, e.g. by some distance measure, but classified to the desired class 1.

**Definition 1.** (CE) Given a NN  $\mathcal{M}_\Theta$ , an input  $x \in \mathcal{X}$  such that  $\mathcal{M}_\Theta(x) < 0$ , and a distance metric  $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ , a CE  $x' \in \mathcal{X}$  is such that:

$$\begin{aligned} & \arg \min_{x'} \text{dist}(x, x') \\ & \text{subject to } \mathcal{M}_\Theta(x') \geq 0 \end{aligned}$$

The minimum distance objective targets the minimum effort by the end user to achieve a change, which corresponds to the basic requirement of proximity mentioned in Section 1. In the literature, normalised  $L_1$  distance is often adopted as the distance metric because it induces changes in fewer features in the CE [Wachter *et al.*, 2017]. However, methods that find such plain CEs usually result in unrealistic combinations of features, or outliers to the underlying data distribution of the training dataset. A plausible CE avoids these issues and is formally defined as follows:

**Definition 2.** (Plausible CE) Given a NN  $\mathcal{M}_\Theta$  and an input  $x \in \mathcal{X}$  such that  $\mathcal{M}_\Theta(x) < 0$ , a distance metric  $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$  and some plausible region  $\mathcal{X}_{\text{plaus}} \subseteq \mathbb{R}^d$ , a plausible CE is an  $x'$  such that:

$$\begin{aligned} & \arg \min_{x'} \text{dist}(x, x') \\ & \text{subject to } \mathcal{M}_\Theta(x') \geq 0, \quad x' \in \mathcal{X}_{\text{plaus}} \end{aligned}$$

The plausible region  $\mathcal{X}_{\text{plaus}}$  may be used to eliminate any unrealistic feature values (e.g. a value of 0.95 for a discrete feature), or to indicate a densely populated region that is close to the data manifold of the training dataset. Additionally, it may also include some actionability considerations, such as restricting immutable attributes (e.g. avoiding suggesting changes in gender) or specifying some relations between input features (e.g. obtaining a doctoral degree should also cost the user at least 4 years).

**Robustness of Counterfactual Explanations.** Studies have shown that CEs found by the above formulations are readily invalidated when small changes occur in the model parameters of the NNs. We formalise this in the following and begin by introducing a distance measure between two NNs and a definition of model shift. Note that Definitions 3 to 7 are adapted from [Jiang *et al.*, 2023].

**Definition 3.** (Distance between two NNs) Consider two NNs  $\mathcal{M}_\Theta, \mathcal{M}_{\Theta'}$  of the same architecture characterised by parameters  $\Theta$  and  $\Theta'$ . For  $0 \leq p \leq \infty$ , the p-distance between  $\mathcal{M}_\Theta$  and  $\mathcal{M}_{\Theta'}$  is  $d_p(\mathcal{M}_\Theta, \mathcal{M}_{\Theta'}) = \|\Theta - \Theta'\|_p$ .

**Definition 4.** (Bounded model shifts) Given a NN  $\mathcal{M}_\Theta$ ,  $\delta \in \mathbb{R}_{>0}$  and  $0 \leq p \leq \infty$ , the set of bounded model shifts is defined as  $\Delta = \{\mathcal{M}_{\Theta'} \mid d_p(\mathcal{M}_\Theta, \mathcal{M}_{\Theta'}) \leq \delta\}$ .

**Certifying Robustness.** Having presented the definitions required to formalise the optimisation problem for finding provably robust and plausible CEs, we now introduce another relevant technique that uses interval abstractions to certify the robustness of CEs. We refer to the certification

process as the  $\Delta$ -robustness test; this will be used for parts of our method and also as an evaluation metric in the experiments. We assume  $p = \infty$  for bounded model shifts  $\Delta$  throughout the paper.

**Definition 5.** (Interval abstraction of NN) Consider a NN  $\mathcal{M}_\Theta$  with  $\Theta = [\theta_0, \dots, \theta_d]$ . Given a set of bounded model shifts  $\Delta$ , we define the interval abstraction of  $\mathcal{M}_\Theta$  under  $\Delta$  as the model  $\mathcal{I}_{(\Theta, \Delta)} : \mathcal{X} \rightarrow \wp\mathbb{R}$  (for  $\wp\mathbb{R}$  the set of all closed intervals over  $\mathbb{R}$ ) such that:

- $\mathcal{M}_\Theta$  and  $\mathcal{I}_{(\Theta, \Delta)}$  have the same architecture;
- $\mathcal{I}_{(\Theta, \Delta)}$  is parameterised by an interval-valued vector  $\Theta = [\theta_0, \dots, \theta_d]$  such that, for  $i \in \{0, \dots, d\}$ ,  $\theta_i = [\theta_i - \delta, \theta_i + \delta]$ , where  $\delta$  is the bound in  $\Delta$ .

When  $p = \infty$ ,  $\theta_i$  encodes the range of possible model parameter changes by the application of  $\Delta$  to  $\mathcal{M}_\Theta$ . Given a fixed input, by propagating the weight and bias intervals, the output range of  $\mathcal{I}_{(\Theta, \Delta)}$  exactly represents the possible output range for the input by applying  $\Delta$  to  $\mathcal{M}_\Theta$  [Jiang *et al.*, 2023].

**Definition 6.** (Interval abstraction of NN classification) Let  $\mathcal{I}_{(\Theta, \Delta)}$  be the interval abstraction of a NN  $\mathcal{M}_\Theta$  under  $\Delta$ . Given an input  $x \in \mathcal{X}$ , let  $\mathcal{I}_{(\Theta, \Delta)}(x) = [l, u]$ . Then, we say that  $\mathcal{I}_{(\Theta, \Delta)}$  classifies  $x$  as class 1 if  $l \geq 0$  (denoted, with an abuse of notation,  $\mathcal{I}_{(\Theta, \Delta)}(x) \geq 0$ ), and as class 0 if  $u < 0$  (denoted, with an abuse of notation,  $\mathcal{I}_{(\Theta, \Delta)}(x) < 0$ ).

Indeed, for an input, if the lower bound  $l$  of pre-sigmoid output node interval  $[l, u]$  of  $\mathcal{I}_{(\Theta, \Delta)}$  satisfies  $l \geq 0$ , then it means all shifted models in  $\Delta$  would predict the input with a pre-sigmoid value that is greater than or equal to 0, all resulting in predicted label 1. We apply this intuition to the CE context:

**Definition 7.** ( $\Delta$ -robust CE) Consider an input  $x \in \mathcal{X}$  and a model  $\mathcal{M}_\Theta$  such that  $\mathcal{M}_\Theta(x) < 0$ . Let  $\mathcal{I}_{(\Theta, \Delta)}$  be the interval abstraction of  $\mathcal{M}_\Theta$  under  $\Delta$ . We say that a CE  $x'$  is  $\Delta$ -robust iff  $\mathcal{I}_{(\Theta, \Delta)}(x') \geq 0$ .

Checking whether a CE  $x'$  is  $\Delta$ -robust requires the calculation of the lower bound  $l$  of the pre-sigmoid output node interval  $[l, u]$  of  $\mathcal{I}_{(\Theta, \Delta)}$ . This process can be encoded as a MILP program (see Appendix B in [Jiang *et al.*, 2023]).

**Optimising for Robustness and Plausibility.** Now we introduce the targeted provably robust and plausible optimisation problem based on Definitions 2 and 7, by taking inspiration from the robust optimisation technique [Ben-Tal *et al.*, 2009].

**Definition 8.** (Provably robust and plausible CE) Given a NN  $\mathcal{M}_\Theta$ , an input  $x \in \mathcal{X}$  such that  $\mathcal{M}_\Theta(x) < 0$ , a distance metric  $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$  and some plausible region  $\mathcal{X}_{\text{plaus}} \subseteq \mathbb{R}^d$ , let  $\mathcal{I}_{(\Theta, \Delta)}$  be the interval abstraction of  $\mathcal{M}_\Theta$  under the bounded model shifts  $\Delta$ . Then, a provably robust and plausible CE  $x' \in \mathcal{X}$  is such that:

$$\arg \min_{x'} \text{dist}(x, x') \tag{1a}$$

$$\text{subject to } \mathcal{I}_{(\Theta, \Delta)}(x') \geq 0, \tag{1b}$$

$$x' \in \mathcal{X}_{\text{plaus}} \tag{1c}$$

The optimisation problem (1) can be equivalently rewritten as follows:

$$\arg \min_{x'} \text{dist}(x, x') \quad (2a)$$

$$\text{subject to } \max_{\mathcal{M}_{\Theta'} \in \Delta} [-\mathcal{M}_{\Theta'}(x')] \leq 0, \quad (2b)$$

$$x' \in \mathcal{X}_{\text{plaus}} \quad (2c)$$

We show next a novel approach for solving this robust optimisation problem (2).

## 4 PROPLACE

The procedure for computing robust and plausible CEs, solving the optimisation problem (2), is summarised in Algorithm 1. We will first introduce how the plausible region  $\mathcal{X}_{\text{plaus}}$  is constructed in Section 4.1 (corresponding to Line 3, Algorithm 1). Then, in Section 4.2 we will present the bi-level optimisation method (corresponding to Lines 4-5, Algorithm 1) to solve the robust optimisation problem (2). In Section 4.2 we will also instantiate the complete bi-level optimisation formulations (in MILP form) of our method for NNs with ReLU activation functions. Finally, in Section 4.3 we discuss the soundness and completeness of Algorithm 1 and prove its convergence.

### 4.1 Identifying Search Space $\mathcal{X}_{\text{plaus}}$

As mentioned in Section 2, points from the training dataset (especially  $k$ -nearest-neighbours) are frequently utilised in the literature to induce plausibility. In this work, we propose to use a more specialised kind of dataset point,  $k$   $\Delta$ -robust nearest-neighbours, to construct the search space for CEs that is both plausible and robust.

**Definition 9.** ( $k$   $\Delta$ -robust nearest-neighbours) *Given a NN  $\mathcal{M}_{\Theta}$  and an input  $x \in \mathcal{X}$  such that  $\mathcal{M}_{\Theta}(x) < 0$ , a distance metric  $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ , a dataset  $\mathcal{D} \subseteq \mathbb{R}^d$  on which  $\mathcal{M}_{\Theta}$  is trained, and a set of bounded model shifts of interest  $\Delta$ , let  $\mathcal{I}_{(\Theta, \Delta)}$  be the interval abstraction of  $\mathcal{M}_{\Theta}$  under  $\Delta$ . Then, the  $k$   $\Delta$ -robust nearest-neighbours of  $x$  is a set  $S_{k, \Delta} \subseteq \mathcal{D}$  with cardinality  $|S_{k, \Delta}| = k$  such that:*

- $\forall x' \in S_{k, \Delta}, x'$  is  $\Delta$ -robust, i.e.  $\mathcal{I}_{(\Theta, \Delta)}(x') \geq 0$ ,
- $\forall x'' \in \mathcal{D} \setminus S_{k, \Delta}$ , if  $x''$  is  $\Delta$ -robust,  $\text{dist}(x, x'') \geq \max_{x' \in S_{k, \Delta}} \text{dist}(x, x')$ .

---

#### Algorithm 1 PROPLACE

---

**Require:** input  $x$ , model  $\mathcal{M}_{\Theta}$ ,

- 1: training dataset  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ ,
  - 2: set of bounded model shifts  $\Delta$ ,
  - 3: plausible region to be used as CEs search space  $\mathcal{X}_{\text{plaus}}$ .
  - 4: **Init:**  $x' \leftarrow \emptyset; \Delta' \leftarrow \{\mathcal{M}_{\Theta}\}$
  - 5: **Repeat until**  $(-\mathcal{M}_{\Theta'}(x')) \leq 0$ 
    - $x' \leftarrow \text{Outer\_minimisation}(\mathcal{M}_{\Theta}, x, \mathcal{X}_{\text{plaus}}, \Delta')$
    - $\mathcal{M}_{\Theta'} \leftarrow \text{Inner\_maximisation}(x', \Delta', \Delta)$
    - $\Delta' \leftarrow \Delta' \cup \{\mathcal{M}_{\Theta'}\}$
  - 6: **return**  $x'$
- 

The first constraint enforces the  $\Delta$ -robustness, and the second states that the points contained in the set are the  $k$  nearest points to the input  $x$  amongst all the  $\Delta$ -robust dataset points. In practice, in order to compute the  $k$   $\Delta$ -robust nearest-neighbours, we fit a  $k$ -d tree on the dataset points that are classified to the desired class, then iteratively query the  $k$ -d tree for the nearest neighbour of an input, until the result satisfies the  $\Delta$ -robustness test (Definition 7).

Restricting the CE search space within the convex hull of these robust neighbours will likely induce high plausibility (and robustness). However, because these points are deep within parts of the training dataset that are classified to another class, they may be far from the model's decision boundary, therefore resulting in large distances to the inputs. In fact, [Dutta *et al.*, 2022; Hamman *et al.*, 2023] adopted similar robust nearest-neighbours (using other notions of robustness tests) as the final CEs, and poor proximity was observed in their experiment results. They have also shown that finding CEs using line search between proximal CEs and these robust neighbours can slightly improve proximity.

In our case, since the validity of the CEs can be guaranteed from the optimisation procedures (Section 4.2), we expand the plausible search space across the decision boundary by taking the input into consideration, which is assumed to also be inside the data distribution.

**Definition 10.** (Plausible region) *Given an input  $x \in \mathbb{R}^d$  and its  $k$   $\Delta$ -robust nearest neighbours  $S_{k, \Delta}$ , the plausible region  $\mathcal{X}_{\text{plaus}}$  is the convex hull of  $S_{k, \Delta} \cup \{x\}$ .*

By restricting the CE search space to such convex hull, the method has the flexibility to find close CEs (with  $x$  as a vertex), or robust and plausible CEs (with the robust neighbours as other vertices). This  $\mathcal{X}_{\text{plaus}}$  ensures the soundness and completeness of our method (Section 4.3).

### 4.2 Bi-level Optimisation Method with MILP

#### 4.2.1 Outer and Inner Optimisation problems

We separate the robust optimisation problem (2) to solve into outer minimisation and inner maximisation problems, as specified in Definitions 11 and 12.

**Definition 11.** *Given a NN  $\mathcal{M}_{\Theta}$  and an input  $x \in \mathcal{X}$  such that  $\mathcal{M}_{\Theta}(x) < 0$ , a distance metric  $\text{dist} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$  and some plausible region  $\mathcal{X}_{\text{plaus}} \subseteq \mathbb{R}^d$ , let  $\Delta'$  be a set of shifted models. Then, the outer minimisation problem finds a CE  $x'$  such that:*

$$\arg \min_{x'} \text{dist}(x, x') \quad (3a)$$

$$\text{subject to } -\mathcal{M}_{\Theta'}(x') \leq 0, \text{ for each } \mathcal{M}_{\Theta'} \in \Delta', \quad (3b)$$

$$x' \in \mathcal{X}_{\text{plaus}} \quad (3c)$$

**Definition 12.** *Given a CE  $x' \in \mathcal{X}$  found by the outer minimisation problem, the set of bounded model shifts  $\Delta$ , the inner maximisation problem finds a shifted model  $\mathcal{M}_{\Theta'}$  such that:*

$$\arg \max_{\mathcal{M}_{\Theta'}} -\mathcal{M}_{\Theta'}(x') \quad (4a)$$

$$\text{subject to } \mathcal{M}_{\Theta'} \in \Delta \quad (4b)$$

The outer minimisation problem relaxes the constraint that a CE should be robust to all possible model shifts in the set  $\Delta$ ; instead, it requires robustness wrt a subset of the model changes  $\Delta' \subset \Delta$ .  $\Delta'$  is initialised with the original classification model  $\mathcal{M}_\Theta$ . At the first execution, the outer minimisation finds the closest CE  $x'$  valid for that model. Then,  $x'$  is passed to the inner maximisation problem to compute the model shift  $S(\mathcal{M}_\Theta)$  that produces the lowest model output score. This model shift is considered to be the worst-case perturbation on the model parameters in the set  $\Delta$ , and is added to  $\Delta'$ . In the next iterations,  $x'$  is updated to the closest CE valid for all the models in  $\Delta'$  (outer), which is being expanded (inner), until convergence.

#### 4.2.2 MILP Formulations

The proposed bi-level optimisation method in Section 4.2.1 is independent of specific NN structures. In this section, we take NNs with ReLU activation functions as an example to further elaborate the method. We denote the total number of hidden layers in an NN  $\mathcal{M}_\Theta$  as  $h$ . We call  $N_0$ ,  $N_{h+1}$ , and  $N_i$  the sets of input, output, and hidden layer nodes for  $i \in [h]$ , and their node values are  $V_0$ ,  $V_{h+1}$ , and  $V_i$ . For hidden layer nodes  $V_i = \text{ReLU}(W_i V_{i-1} + B_i)$ , and for output layer nodes  $V_{h+1} = W_{h+1} V_h + B_{h+1}$ , where  $W_i$  is the weight matrix connecting nodes at layers  $i-1$  and  $i$ , and  $B_i$  is the bias vector of nodes  $N_i$ . We instantiate the formulations using normalised  $L_1$ , while our method PROPLACE can accommodate arbitrary distance metrics.

The outer minimisation problem is equivalent to the following MILP program, where the superscripts  $j$  on weight matrices and bias vectors indicate they are model parameters of the  $j$ -th model  $\mathcal{M}_\Theta^j \in \Delta'$ :

$$\min_{x', y, \lambda} \|x - x'\|_1 \quad (5a)$$

$$\text{s.t. } V_0^j = x', \quad (5b)$$

$$y_i^j \in \{0, 1\}^{|N_i|}, i \in [h], j \in [\Delta'] \quad (5c)$$

$$0 \leq V_i^j \leq M y_i^j, i \in [h], j \in [\Delta'] \quad (5d)$$

$$W_i^j V_{i-1}^j + B_i^j \leq V_i^j \leq (W_i^j V_{i-1}^j + B_i^j) + M(1 - y_i^j), \quad (5e)$$

$$i \in [h], j \in [\Delta']$$

$$W_{h+1}^j V_h^j + B_{h+1}^j \geq 0, j \in [\Delta'] \quad (5f)$$

$$\lambda_l \in [0, 1], l \in [|S_{k,\Delta} \cup \{x'\}|], \sum_{l=1}^{|S_{k,\Delta} \cup \{x'\}|} \lambda_l = 1$$

$$x' = \sum_{l=1}^{|S_{k,\Delta} \cup \{x'\}|} \lambda_l x'_l, x'_l \in S_{k,\Delta} \cup \{x\} \quad (5g)$$

Constraints (5b) - (5f) and constraint (5g) correspond respectively to the robustness and plausibility requirement in (3b) - (3c).

The inner maximisation program can be formulated as the following MILP program, where the superscripts 0 on weight matrices and biases indicate they are model parameters of the original model  $\mathcal{M}_\Theta$ , and  $\delta$  is the bound of model magni-

tude change specified in  $\Delta$ :

$$\max_{W, B, y} V_{h+1} \quad (6a)$$

$$\text{s.t. } V_0 = x', \quad (6b)$$

$$y_i \in \{0, 1\}^{|N_i|}, i \in [h] \quad (6c)$$

$$0 \leq V_i \leq M y_i, i \in [h] \quad (6d)$$

$$W_i V_{i-1} + B_i \leq V_i \leq (W_i V_{i-1} + B_i) + M(1 - y_i), \quad (6e)$$

$$i \in [h]$$

$$V_{h+1} = W_{h+1} V_h + B_{h+1} \quad (6f)$$

$$W_i^0 - \delta \leq W_i \leq W_i^0 + \delta, i \in [h+1] \quad (6g)$$

$$B_i^0 - \delta \leq B_i \leq B_i^0 + \delta, i \in [h+1] \quad (6h)$$

Due to the flexibility of such MILP programs, the framework accommodates continuous, ordinal, and categorical features [Mohammadi *et al.*, 2021]. Specific requirements like feature immutability or associations between features can also be encoded [Ustun *et al.*, 2019]. These MILP problems can be directly solved using off-the-shelf solvers such as Gurobi [Gurobi Optimization, LLC, 2023].

#### 4.3 Soundness, Completeness and Convergence of Algorithm 1

We now discuss the soundness and completeness of our method by restricting the search space for the CE to the plausible region  $\mathcal{X}_{\text{plaus}}$ . From its definition, the vertices (except the input  $x$ ) of  $\mathcal{X}_{\text{plaus}}$  are  $\Delta$ -robust, which thus satisfies the robustness requirement of our target problem (Definition 8). This means that there exist at least  $k$  points in the search space satisfying constraint (2c) that also satisfy constraint (2b), making these points feasible solutions for the target problem. We may thus make the following remark:

**Proposition 1.** *Algorithm 1 is sound and complete if  $\exists x' \in \mathcal{D}$  such that  $x'$  is  $\Delta$ -robust.*

Next, we adapt the method in [Mutapcic and Boyd, 2009, Section 5.2] to provide an upper bound on the maximum number of iterations of Algorithm 1.

**Proposition 2.** *Given the requirements of Algorithm 1, assume the classifier  $\mathcal{M}_\Theta$  is Lipschitz continuous in  $x'$ . Then, the maximum number of iterations before Algorithm 1 terminates is bounded.*

*Proof.* Firstly, we assume two small tolerance variables  $\sigma > t > 0$  and modify the robustness constraint (2b) of Definition 8 to:  $\max_{\mathcal{M}_{\Theta'} \in \Delta} [-\mathcal{M}_{\Theta'}(x') + \sigma] \leq t$ , such that the cor-

rectness of the robustness guarantee is not affected. The termination condition for Algorithm 1 therefore becomes  $-\mathcal{M}_{\Theta'}(x') + \sigma \leq t$ .

Consider the plausible CE problem (Definition 2 with the validity constraint modified to  $-\mathcal{M}_\Theta(x') + \sigma \leq t$ ), which is the problem solved by the first execution (iteration 1) of the outer minimisation problem in Algorithm 1. We denote its feasible region as  $\mathcal{F}$ . Suppose  $\mathcal{M}_\Theta$  is a ReLU NN without the final (softmax or sigmoid) activation layer, then  $\mathcal{M}_\Theta$  is Lipschitz continuous. Let  $f(x', \mathcal{M}_\Theta) := -\mathcal{M}_\Theta(x') + \sigma$ , then  $f$  is Lipschitz continuous in  $x'$  over  $\mathcal{F}$  with some Lipschitz

constant  $L$ . For a distance metric  $dist : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ , and any  $x_1, x_2 \in \mathcal{F}$ , we have:

$$|f(x_1, \mathcal{M}_\Theta) - f(x_2, \mathcal{M}_\Theta)| \leq L \times dist(x_1, x_2) \quad (7)$$

At iteration  $m$ , we denote the CE found by the outer minimisation as  $x'^{(m)}$ , and the shifted model found by the inner maximisation as  $\mathcal{M}_\Theta^{(m)}$ . Then,  $f(x'^{(m)}, \mathcal{M}_\Theta^{(m)}) := -\mathcal{M}_\Theta^{(m)}(x'^{(m)}) + \sigma$ . Assume at step  $m$  the algorithm has not terminated, then

$$f(x'^{(m)}, \mathcal{M}_\Theta^{(m)}) > t \quad (8)$$

For the iteration steps  $n > m$ ,  $x'^{(n)}$  is required to be valid on  $\mathcal{M}_\Theta^{(m)}$  as specified in the outer minimisation problem, we therefore have:

$$f(x'^{(n)}, \mathcal{M}_\Theta^{(m)}) \leq 0 \quad (9)$$

Combining (8) and (9) yields

$$f(x'^{(m)}, \mathcal{M}_\Theta^{(m)}) - f(x'^{(n)}, \mathcal{M}_\Theta^{(m)}) > t \quad (10)$$

Further combining (10) with (7), for the iteration steps  $n > m$ ,

$$dist(x'^{(m)}, x'^{(n)}) > \frac{t}{L} \quad (11)$$

Consider the balls  $B_i, i = 1, \dots, m$ , of diameter  $\frac{t}{L}$  centred at each intermediate result of the outer minimisation problem,  $x'^{(i)}$ . From (11), it can be concluded that for any two intermediate  $x'^{(i)}, x'^{(j)}, 1 < i, j < m$ ,  $dist(x'^{(i)}, x'^{(j)}) > \frac{t}{L}$ , and  $x'^{(i)}$  and  $x'^{(j)}$  are the centres of the balls  $B^{(i)}$  and  $B^{(j)}$ . Therefore, any two circles will not intercept. The total volume of these balls is thus  $m \times U \times (\frac{t}{L})^d$ , where  $U$  is the unit volume in  $\mathbb{R}^d$ .

Consider a ball that encompasses the feasible solution region  $\mathcal{F}$  and let  $R$  be its radius. We know that  $x'_i, i = 1, \dots, m$ , are all within the feasible region  $\mathcal{F}$ , therefore, the ball  $B$  that has a radius  $R + \frac{t}{2L}$  will cover the spaces of the small balls  $B_i, i = 1, \dots, m$ . Also, the volume of  $B$  is  $U \times (2R + \frac{t}{L})^d$  and will be greater than the total volume of the small balls, which means:

$$U \times \left(2R + \frac{t}{L}\right)^d > m \times U \times \left(\frac{t}{L}\right)^d \implies m < \left(\frac{2RL}{t} + 1\right)^d$$

It can be concluded that the step number at which Algorithm 1 has not terminated is bounded above by the  $(\frac{2RL}{t} + 1)^d$ .  $\square$

## 5 Experiments

In this section, we demonstrate that our proposed method achieves state-of-the-art performances compared with existing robust CEs generation methods.

**Datasets and Classifiers.** Our experiments use four benchmark datasets in financial and legal contexts: the Adult Income (ADULT), COMPAS, Give Me Some Credits (GMC), and HELOC datasets. We adopt the pre-processed versions available in the CARLA library [Pawelczyk *et al.*, 2021] where each dataset contains binarised categorical features and min-max scaled continuous features. Labels 0 and 1 are

the unwanted and the desired class, respectively. We split each dataset into two halves. We use the first half for training NNs with which the robust CEs are generated, and the second half for model retraining and evaluating the robustness of the CEs.

For making predictions and generating CEs, the NNs contain two hidden layers with ReLU activation functions. They are trained using the Adam optimiser with a batch size of 32, and under the standard 80%, 20% train-test dataset split setting. The classifiers achieved 84%, 85%, 94%, and 76% accuracies on the test set of ADULT, COMPAS, GMC, and HELOC datasets, respectively.

The retrained models have the same hyperparameters and training procedures as the original classifiers. Following the experimental setup in previous works [Dutta *et al.*, 2022; Ferrario and Loi, 2022; Nguyen *et al.*, 2022; Black *et al.*, 2022; Upadhyay *et al.*, 2021], for each dataset, we train 10 new models using both halves of the dataset to simulate the possible retrained models after new data are collected. We also train 10 new models using 99% of the first half of the dataset (different 1% data are discarded for each training), to simulate the leave-one-out retraining procedures. The random seed is perturbed for retraining. These 20 retrained models are used for evaluating the robustness of CEs.

**Evaluation Metrics.** The CEs are evaluated by the following metrics for their proximity, plausibility, and robustness.

- $\ell_1$  measures the average  $L_1$  distance between a CE and its corresponding input.
- $lof$  is the average 10-Local Outlier Factor [Breunig *et al.*, 2000] of the generated CEs, which indicates to what extent a data point is an outlier wrt its  $k$  nearest neighbours in a specified dataset.  $lof$  values close to 1 indicate inliers, larger values (especially if greater than 1.5) indicate outliers.
- $vr$ , the validity of CEs on the retrained models, is defined as the average percentage of CEs that remain valid (classified to class 1) under the retrained models.
- $v\Delta$  is the percentage of CEs that are  $\Delta$ -robust. The bound of model parameter changes  $\delta$  is specified to be the same as the value used in our algorithm.

**Baselines.** We compare our method with six state-of-the-art methods for generating CEs, including five which target robustness. WCE [Wachter *et al.*, 2017] is the first method to generate CEs for NNs, which minimises the  $\ell_1$  distance between the CEs and the inputs. Robust Bayesian Recourse (RBR) [Nguyen *et al.*, 2022] addresses the proximity, robustness, and plausibility of CEs. RobXNN [Dutta *et al.*, 2022] is a nearest-neighbour-based method that focuses on a different notion of robustness to model changes. Robust Algorithmic Recourse (ROAR) [Upadhyay *et al.*, 2021] optimises for proximity and the same  $\Delta$  notion of robustness. Proto-R and MILP-R are the methods proposed by [Jiang *et al.*, 2023] which embed the  $\Delta$ -robustness test into the base methods of [Van Looveren and Klaise, 2021] and [Mohammadi *et al.*, 2021]. For all methods including ours, we tune their hyperparameters to maximise the validity after retraining  $vr$ .

|             | ADULT         |                       |                       |                  | COMPAS        |                       |                       |                  | GMC           |                       |                       |                  | HELOC         |                       |                       |                  |
|-------------|---------------|-----------------------|-----------------------|------------------|---------------|-----------------------|-----------------------|------------------|---------------|-----------------------|-----------------------|------------------|---------------|-----------------------|-----------------------|------------------|
|             | vr $\uparrow$ | v $\Delta$ $\uparrow$ | $\ell_1$ $\downarrow$ | lof $\downarrow$ | vr $\uparrow$ | v $\Delta$ $\uparrow$ | $\ell_1$ $\downarrow$ | lof $\downarrow$ | vr $\uparrow$ | v $\Delta$ $\uparrow$ | $\ell_1$ $\downarrow$ | lof $\downarrow$ | vr $\uparrow$ | v $\Delta$ $\uparrow$ | $\ell_1$ $\downarrow$ | lof $\downarrow$ |
| WCE         | 89            | 78                    | .175                  | 1.59             | 57            | 0                     | .170                  | 1.81             | 84            | 18                    | .148                  | 2.80             | 49            | 0                     | .045                  | 1.16             |
| RBR         | 100           | 0                     | .031                  | 1.28             | 100           | 62                    | .043                  | 1.34             | 90            | 0                     | .050                  | 1.31             | 80            | 0                     | .038                  | 1.10             |
| RobXNN      | 100           | 82                    | .064                  | 1.28             | 100           | 82                    | .050                  | 1.11             | 100           | 96                    | .073                  | 1.35             | 100           | 30                    | .073                  | 1.04             |
| ROAR        | 99            | 98                    | .279                  | 1.96             | 98            | 100                   | .219                  | 2.84             | 96            | 88                    | .188                  | 4.22             | 98            | 98                    | .109                  | 1.57             |
| PROTO-R     | 98            | 55                    | .068                  | 1.60             | 100           | 100                   | .084                  | 1.36             | 100           | 100                   | .066                  | 1.49             | 100           | 100                   | .057                  | 1.21             |
| MILP-R      | 100           | 100                   | .024                  | 1.69             | 100           | 100                   | .040                  | 1.71             | 100           | 100                   | .059                  | 2.08             | 100           | 100                   | .044                  | 2.48             |
| <b>OURS</b> | 100           | 100                   | .046                  | 1.22             | 100           | 100                   | .039                  | 1.24             | 100           | 100                   | .058                  | 1.24             | 100           | 100                   | .057                  | 1.04             |

Table 1: Evaluations of PROPLACE (OURS) and baselines on NNs. The  $\uparrow$  ( $\downarrow$ ) indicates that higher (lower) values are preferred for the evaluation metric.

**Results.** We randomly select 50 test points from each dataset that are classified to be the unwanted class, then apply our method and each baseline to generate CEs for these test points. Results are shown in Table 1.

As a non-robust baseline, the WCE method is the least robust while producing high  $\ell_1$  costs and poor plausibility. Though RBR shows the lowest  $\ell_1$  results on three datasets, it has only moderate robustness against the naturally retrained models and is not  $\Delta$ -robust on any dataset. The rest of the baselines all show strong robustness on at least three datasets, with our method having slightly better  $vr$  and  $v\Delta$  results, evaluated at 100% in every experiment. This indicates that our method PROPLACE can not only guarantee robustness under bounded model parameter changes but also induce reliable robustness against unbounded model changes. In terms of plausibility, our method shows the best lof score in most experiments. Therefore, our method has addressed the limitation in the literature that no method optimises for guaranteed  $\Delta$ -robustness and plausibility. Though the two properties have established trade-offs with proximity [Pawelczyk *et al.*, 2020b; Pawelczyk *et al.*, 2022; Upadhyay *et al.*, 2021], our method still shows  $\ell_1$  costs lower than all methods except RBR, which is significantly less robust, and MILP-R, which finds outliers. For the COMPAS dataset, our method has the best proximity result among all baselines.

Note that the PROTO-R baseline from the work which proposed certification for  $\Delta$ -robustness failed to find  $\Delta$ -robust CEs on the ADULT dataset, as was the case in their results (see Table 1, [Jiang *et al.*, 2023]). This is due to the fact that their method rely heavily on a base method to find CEs, and it is not straightforward to be always able to direct the hyperparameters search for optimising  $\Delta$ -robustness. With improved soundness and completeness (Proposition 1), PROPLACE always finds provably robust results.

## 6 Conclusions

We proposed a robust optimisation framework PROPLACE to generate provably robust and plausible CEs for neural networks. The method addresses the limitation in the literature that existing methods lack formal robustness guarantees to bounded model parameter changes and do not generate plausible CEs. We proved the soundness, completeness, and convergence of PROPLACE. Through a comparative study, we show the efficacy of our method, demonstrating the best ro-

bustness and plausibility results with better proximity than the most robust baselines. Despite the specific form of robustness we target, PROPLACE is also empirically robust to model retraining with unbounded parameter changes. Future work could include investigating the properties of actionability and diversity, evaluations with user studies, and investigating connections between  $\Delta$ -robustness and different notions of robustness measures.

## Acknowledgement

Jiang, Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Jianglin Lan is supported by a Leverhulme Trust Early Career Fellowship under Award ECF-2021-517. Leofante is supported by an Imperial College Research Fellowship grant. Rago and Toni were partially funded by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 101020934). Any views or opinions expressed herein are solely those of the authors listed.

## References

- [Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Adv. in Neural Information Processing Systems 31, NeurIPS*, 2018.
- [Bajaj *et al.*, 2021] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Adv. in Neural Information Processing Systems 34, NeurIPS*, 2021.
- [Ben-Tal *et al.*, 2009] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.
- [Black *et al.*, 2022] Emily Black, Zifan Wang, and Matt Fredrikson. Consistent counterfactuals for deep models. In *10th Int. Conf. on Learning Representations, ICLR*, 2022.
- [Breunig *et al.*, 2000] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *ACM SIGMOD Int. Conf. on Management of Data*, 2000.

- [Bui *et al.*, 2022] Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. Counterfactual plans under distributional ambiguity. In *10th Int. Conf. on Learning Representations, ICLR*, 2022.
- [Celar and Byrne, 2023] Lenart Celar and Ruth M. J. Byrne. How people reason with counterfactual and causal explanations for AI decisions in familiar and unfamiliar domains. *Memory & Cognition*, 2023.
- [Dhurandhar *et al.*, 2018] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Adv. in Neural Information Processing Systems 31, NeurIPS*, 2018.
- [Dominguez-Olmedo *et al.*, 2022] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *39th Int. Conf. on Machine Learning, ICML*, 2022.
- [Dutta *et al.*, 2022] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *39th Int. Conf. on Machine Learning, ICML*, 2022.
- [Ferrario and Loi, 2022] Andrea Ferrario and Michele Loi. The robustness of counterfactual explanations over time. *IEEE Access*, 10:82736–82750, 2022.
- [Guidotti, 2022] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.
- [Gurobi Optimization, LLC, 2023] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.
- [Hamman *et al.*, 2023] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *40th Int. Conf. on Machine Learning, ICML*, 2023.
- [Huai *et al.*, 2022] Mengdi Huai, Jinduo Liu, Chenglin Miao, Liuyi Yao, and Aidong Zhang. Towards automating model explanations with certified robustness guarantees. In *36th AAAI Conf. on Artificial Intelligence, AAAI*, 2022.
- [Jiang *et al.*, 2023] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *37th AAAI Conf. on Artificial Intelligence, AAAI*, 2023.
- [Leofante and Lomuscio, 2023a] Francesco Leofante and Alessio Lomuscio. Robust explanations for human-neural multi-agent systems with formal verification. In *20th Eur. Conf. on Multi-Agent Systems EUMAS*, 2023.
- [Leofante and Lomuscio, 2023b] Francesco Leofante and Alessio Lomuscio. Towards robust contrastive explanations for human-neural multi-agent systems. In *22nd Int. Conf. on Autonomous Agents and Multiagent Systems, AAMAS*, 2023.
- [Leofante *et al.*, 2023] F Leofante, E Botoeva, and V Rajani. Counterfactual explanations and model multiplicity: a relational verification view. In *The 20th Int. Conf. on Principles of Knowledge Representation and Reasoning, KR*, 2023.
- [Maragno *et al.*, 2023] Donato Maragno, Jannis Kurtz, Tabea E Röber, Rob Goedhart, Ş Ilker Birbil, and Dick den Hertog. Finding regions of counterfactual explanations via robust optimization. *arXiv:2301.11113*, 2023.
- [Marques-Silva and Ignatiev, 2022] João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *36th AAAI Conf. on Artificial Intelligence, AAAI*, 2022.
- [Mohammadi *et al.*, 2021] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. Scaling guarantees for nearest counterfactual explanations. In *Conf. on AI, Ethics, and Society, AIES*, 2021.
- [Mothilal *et al.*, 2020] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual expl. In *Conf. on Fairness, Accountability, and Transparency*, 2020.
- [Mutapcic and Boyd, 2009] Almir Mutapcic and Stephen Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.
- [Nguyen *et al.*, 2022] Tuan-Duy H Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Robust bayesian recourse. In *38th Conf. on Uncertainty in AI, UAI*, 2022.
- [Pawelczyk *et al.*, 2020a] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *The ACM Web Conference, WWW*, 2020.
- [Pawelczyk *et al.*, 2020b] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *36th Conf. on Uncertainty in AI, UAI*, 2020.
- [Pawelczyk *et al.*, 2021] 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. *Adv. in Neural Information Processing Systems 34, NeurIPS (Datasets and Benchmarks)*, 2021.
- [Pawelczyk *et al.*, 2022] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. *arXiv:2203.06768*, 2022.
- [Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. FACE: feasible and actionable counterfactual explanations. In *Conf. on AI, Ethics, and Society, AIES*, 2020.
- [Prabhakar and Afzal, 2019] Pavithra Prabhakar and Zahra Rahimi Afzal. Abstraction based output range



- analysis for neural networks. In *Adv. in Neural Information Processing Systems 32, NeurIPS*, 2019.
- [Rawal *et al.*, 2020] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv:2012.11788*, 2020.
- [Sharma *et al.*, 2020] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Conf. on AI, Ethics, and Society, AIES*, 2020.
- [Tolomei *et al.*, 2017] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2017.
- [Upadhyay *et al.*, 2021] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *Adv. in Neural Information Processing Systems 34, NeurIPS*, 2021.
- [Ustun *et al.*, 2019] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Conf. on Fairness, Accountability, and Transparency*, 2019.
- [Van Looveren and Klaise, 2021] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Eur. Conf. on Machine Learning and Knowledge Discovery in Databases, ECML PKDD*, 2021.
- [Verma *et al.*, 2022] Sahil Verma, Keegan Hines, and John P Dickerson. Amortized generation of sequential algorithmic recourses for black-box models. In *36th AAAI Conf. on Artificial Intelligence, AAAI*, 2022.
- [Virgolin and Fracaros, 2023] Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual expl. to adverse perturbations. *Artificial Intelligence*, 316:103840, 2023.
- [Vo *et al.*, 2023] Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. Feature-based learning for diverse and privacy-preserving counterfactual explanations. In *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2023.
- [Wachter *et al.*, 2017] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.
- [Zhang *et al.*, 2023] Songming Zhang, Xiaofeng Chen, Shiping Wen, and Zhongshan Li. Density-based reliable and robust explainer for counterfactual explanation. *Expert Systems with Applications*, pages 120–214, 2023.