Fairness and Robustness of Contrasting Explanations*

André Artelt[1] and Barbara Hammer[1]

CITEC - Cognitive Interaction Technology
Bielefeld University, 33619 Bielefeld, Germany
{aartelt,bhammer}@techfak.uni-bielefeld.de

Abstract. Fairness and explainability are two important and closely related requirements of decision making systems. While fairness and explainability of decision making systems have been extensively studied independently, only little effort has been put into studying fairness of explanations on their own. Current explanations can be unfair to an individual: an example is given by counterfactual explanations which propose different actions to change the output class to two similar individuals. In this work we formally and empirically study individual fairness and its mathematical formalization as robustness for counterfactual explanations as a prominent instance of contrasting explanations. In addition, we propose to use plausible counterfactuals instead of closest counterfactuals for improving the individual fairness of counterfactual explanations.

Keywords: XAI · Contrastive Explanations · Counterfactual Explanations · Fairness · Robustness

1 Introduction

Fairness and transparency are fundamental building blocks of ethical artificial intelligence (AI) and machine learning (ML) based decision making systems. In particular, the increasing use of automated decision making systems have strengthened the demand for trustworthy systems. The criticality of transparency was also recognized by policy makers which resulted in legal regulations like the EU's GDPR [33] that grants the user a right to an explanation. Therefore, explainability and transparency of AI and ML methods has become an important research focus in the last years [16,19,38,41]. Nowadays, there exist many different technologies for explaining ML models [19, 31]. One specific family of methods are model-agnostic methods [19, 35], which are not tailored to a particular model or representation and thus applicable to many different types of ML models.

Examples of model-agnostic methods are feature interaction methods [17], feature importance methods [15], partial dependency plots [46] and methods that approximates the model locally by an explainable model [18, 36]. These methods explain the models by using *features* as vocabulary.

* We gratefully acknowledge funding from the VW-Foundation for the project *IMPACT* funded in the frame of the funding line *AI and its Implications for Future Society*.

A different class of model-agnostic explanations are *example-based explanations* where a prediction or behavior is explained by a (set of) data points [1]. Instances of example-based explanations are prototypes & criticisms [23] and influential instances [24]. Another instance of example-based explanations are counterfactual explanations [43]. A counterfactual explanation is a change of the original input that leads to a different (specific) prediction or behavior of the decision making system - *what has to be different in order to change the prediction of the system?* Such an explanation is considered to be intuitive and useful because it proposes changes to achieve a desired outcome, i.e. it provides *actionable feedback* [31, 43]. There exists strong evidence that explanations by humans are often counterfactual in nature [11]. We will focus on theses types of explanations in this work.

Despite the recent success stories of AI and ML systems, ML systems were also involved in prominent failures like predictive policing [6] and loan approval [44]. Many of those failures, where the systems exhibit unethical behaviour, deal with predictions that are correlated to sensitive attributes like race, gender, etc.

As a consequence, a number of (formal) fairness criteria has been proposed [12, 30]. A large group of criteria are concerned with the dependency between a sensitive attribute/feature and the response variable (prediction/behaviour of the system) - these criteria belong to group-based fairness criteria. Prominent examples are [12] demographic parity, equalized odds, predictive rate parity, or causal independence. However, there do exist fundamental limitations for an efficient implementation of fair models, since some intuitively relevant fairness criteria can be contradictory in specific examples.

Another formalization of fairness focus on individual (rather than group) fairness [14]. The idea behind individual fairness is to treat similar individuals similarly - this notion resembles the concept of individual fairness from other scientific disciplines [9, 14]. Despite its appealing intuition, a major problem of individual fairness constitutes in its formalization – i.e. given two individuals, how can we score their mutual similarity?

In this work we will focus on individual fairness of *explanations* rather than models itself; this question is particularly relevant if explanations lead to actions, such as counterfactual explanations: these can be interpreted as proposed actions in order to achieve a desired goal, hence it is problematic, if the proposed actions differ considerably in between individuals without a clear reason. We will formalize the notion of individual fairness of explanations in terms of a robustness criterion, and we will investigate when and how this criterion can be guaranteed.

Related Work: While fairness and robustness are known to be closely related to each other [13, 32], robustness of explanations has only recently been investigated in first approaches [3,21]. For instance it has been shown that explanation methods are vulnerable to adversarial attacks [3, 5, 22] - i.e. an explanation (such as a saliency map or feature importances) can be (arbitrarily) changed by applying only small perturbations to the original sample which is going to be explained. This instability of explanations also holds true for counterfactual explanations [27]. The necessity of local stability of explanations is widely accepted [4, 5, 22, 27, 28]. However, the challenge how to compute stable and robust counterfactual explanations is still an open-research problem [42].

Fairness and explanations are closely related to each other: For instance, counterfactual explanations can be used for detecting bias and unfairness in decision making systems [39]. Given the complex meaning and definition of fairness, the authors of [40] propose to use explanation methods for explaining the (un-) fairness of a model to a lay person. Explanation methods like counterfactual explanations can also be used for defining fairness criteria such as counterfactual fairness [25] – here a decision making system is considered fair if a change of the sensitive attribute and its induced causal changes, while holding everything else constant, must not change the prediction of the decision making system.

It has been observed that missing stability and robustness can lead to unfair explanations and thus compromise the system's trustworthiness [4, 5]. Only few papers study the fairness of explanations itself, however. For instance the authors of [20] propose a group fairness criteria "Equalizing Recourse" which aims for ensuring that the same opportunities hold for feasible recourse (as suggested by an explanation) across different groups. The authors of [26] propose and study group fairness and individual fairness of causal recourse - i.e. recourse/explanations that are obtained from a given structural cause model (SCM).

Our Contributions: In this work we formalize individual fairness of counterfactual explanations via robustness, and we investigate its formal mathematical properties and its experimental behavior for a few popular models. We propose to use plausible counterfactuals instead of closest counterfactual explanations because we find evidence that the latter yield a better individual fairness.

To address these goal, first, we briefly review counterfactual explanations and individual fairness section 2. Next, we formally define our notion of individual fairness of contrasting explanations in section 3.1. We study formal properties of the fairness and robustness of counterfactual explanations in section 3.3 - we propose the concept to use plausible instead of closest counterfactual explanations for improving the individual fairness of the explanations. We empirically evaluate the individual fairness of closest and plausible counterfactual explanations in section 4. Finally, our work closes with a summary and outlook in section 5.

Note that all proofs and derivations, as well as additional plots of the experiments, can be found in the appendices A and B.

2 Foundations

2.1 Counterfactual Explanations

Counterfactual explanations (often just called counterfactuals) contrast samples by counterparts with minimum change of the appearance but different class label [31, 43] and can be formalized as follows:

Definition 1 (Counterfactual explanation [43]). *Assume a prediction function $h : \mathbb{R}^d \to \mathcal{Y}$ is given. Computing a counterfactual $\boldsymbol{x}' \in \mathbb{R}^d$ for a given input*

$\boldsymbol{x} \in \mathbb{R}^d$ *is phrased as an optimization problem:*

$$\underset{\boldsymbol{x}' \,\in\, \mathbb{R}^d}{\arg\min}\; \ell\left(h(\boldsymbol{x}'), y'\right) + C \cdot \theta(\boldsymbol{x}', \boldsymbol{x}) \tag{1}$$

where $\ell(\cdot)$ *denotes a suitable loss function,* y' *the requested prediction, and* $\theta(\cdot)$ *a penalty term for deviations of* $\boldsymbol{x}'$ *from the original input* $\boldsymbol{x}$*.* $C > 0$ *denotes the regularization strength. We refer to a counterfactual given* $\boldsymbol{x}$ *and desired class* y' *as any solution of the optimization problem.*

In this work we assume that data come from a real-vector space and continuous optimization is possible. In this context [7], two common regularizations are the weighted Manhattan distance and the generalized L_2 distance. Depending on the model and the choice of $\ell(\cdot)$ and $\theta(\cdot)$, the final optimization problem might be differentiable or not. If it is differentiable, we can use a gradient-based optimization algorithm like BFGS. Otherwise, we can rely on black-box optimization algorithms for continuous optimization like Nelder-Mead. Note that counterfactuals need not be unique due to the possible existence of local optima of the optimization problem. We will deal with popular but simple models in the following where this is not the case. In general, we assume that an ordering of counterfactuals is induced by the respective optimization algorithm. Given input $\boldsymbol{x}$ and desired label y', the counterfactual provided by the algorithm at hand is referred to as $\boldsymbol{x}' = \mathrm{CF}(\boldsymbol{x}, y')$.

While the formalization of the optimization problem Eq. (1) is model agnostic (i.e. it does not make any assumptions on the model h), it can be beneficial to rewrite the optimization problem Eq. (1) in constrained form [7]:

$$\underset{\boldsymbol{x}' \,\in\, \mathbb{R}^d}{\arg\min}\; \theta(\boldsymbol{x}', \boldsymbol{x}) \tag{2a}$$

$$\text{s.t. } h(\boldsymbol{x}') = y' \tag{2b}$$

The authors of [7] have shown that the constrained optimization problem Eq. (2) can be turned (or efficiently approximated) into convex programs for many standard machine learning models including generalized linear models, quadratic discriminant analysis, nearest neighbor classifiers, etc. Since convex programs can be solved efficiently [10], the constrained form [7] becomes superior over the original black-box modelling [43] if we have access to the underlying model h.

Counterfactuals stated in its simplest form, like in Definition 1 (also called closest counterfactuals), are similar to adversarial examples, since there are no guarantees that the resulting counterfactual is plausible and feasible in the data domain. As a consequence, the absence of such constraints often leads to counterfactual explanations that are not plausible [8, 29, 34]. To overcome this problem, several approaches [8, 29] propose to allow only those samples that lie on the data manifold - e.g. by enforcing a lower threshold $\delta > 0$ for their probability or density. In particular, the authors of [8] build upon the constrained form Eq. (2) and propose the following extension of Eq. (2) for computing plausible counterfactuals:

$$\underset{\boldsymbol{x}' \in \mathbb{R}^d}{\arg\min}\ \theta(\boldsymbol{x}', \boldsymbol{x}) \tag{3a}$$

$$\text{s.t. } h(\boldsymbol{x}') = y' \tag{3b}$$

$$\hat{p}_y(\boldsymbol{x}') \geq \delta \tag{3c}$$

where $\hat{p}_y(\cdot)$ denotes a class dependent density estimator. Because the true density is usually not known, the density constraint Eq. (3c) can be replaced with an approximation via a Gaussian mixture model (GMM) which then can phrased as a set of convex quadratic constraints [7].

Counterfactuals can be considered unfair to an individual in some settings, since it might happen, that counterfactual explanations are conceptually very different for two similar persons: in particular if $\theta(\cdot)$ is the L_1 norm, explanations might choose entirely different features based on which to explain the decision for neighbored samples. Next, we define the notion of individual fairness.

2.2 Individual Fairness

Individual fairness requires to "treat similar individuals similarly" [14]. Given a prediction function $h : \mathcal{X} \to \mathcal{Y}$, we can formalize individual fairness as follows:

$$\mathrm{d}(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq \epsilon \implies \Delta\big(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)\big) \leq \epsilon \qquad \forall\, \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X} \tag{4}$$

where $\mathrm{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denotes a similarity measure on the individuals in $\mathcal{X}$, $\epsilon > 0$ denotes a threshold up to which we consider two individuals as similar and $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denotes a similarity measure on the predictions of the ML system h. A critical choice, which highly depends on the specific use-case, are the similarity measures $\mathrm{d}(\cdot)$ and $\Delta(\cdot)$. A common choice is a p-norm assuming real-vector spaces. This is also the setting we deal with in this work.

3 Fairness and Robustness of Counterfactual Explanations

We are not interested in individual fairness of the model itself, but the properties of the counterfactual explanations which are provided to explain the model. In the sub sequel we formally study and define individual fairness of explanations via the notion of robustness of counterfactual explanations in general, and investigate its specific properties for important specific prediction functions $h : \mathcal{X} \to \mathcal{Y}$.

3.1 Individual Fairness of Counterfactual Explanations

We aim for a formalization of individual fairness of counterfactual explanations. Inspired by the intuition (and formalization) of individual fairness in section 2.2 (which is similar to the idea of robustness [3]), we propose the following definition that formalizes the intuition that counterfactual explanations of similar individuals should be similar:

Definition 2 (Individual fairness of counterfactual explanations). *Let $h : \mathcal{X} \to \mathcal{Y}$ be a prediction function and $(\boldsymbol{x}_{orig}, y_{orig}) \in \mathcal{X} \times \mathcal{Y}$ with $h(\boldsymbol{x}_{orig}) = y_{orig}$ be a sample prediction that has to be explained. Let $\boldsymbol{x}' = CF(\boldsymbol{x}_{orig}, y') \in \mathcal{X}$ be a counterfactual explanations of $\boldsymbol{x}_{orig}$.*

Let $\boldsymbol{x}$ be a perturbed sample of $\boldsymbol{x}_{orig}$ that is $\mathrm{d}(\boldsymbol{x}_{orig}, \boldsymbol{x}) \leq \epsilon$ for some suitable metric $\mathrm{d}(\cdot)$ and perturbation size ϵ. We denote the according ball of all possible perturbations as $\mathcal{B}_\epsilon(\boldsymbol{x}_{orig})$. Denote $y = y'$ if $h(\boldsymbol{x}) = y_{orig}$ and $y = y_{orig}$ otherwise. Let $\boldsymbol{x}'' = CF(\boldsymbol{x}, y) \in \mathcal{X}$ be a counterfactual explanations of this perturbed sample $(\boldsymbol{x}, y_{orig})$.

We measure the amount of individual (un)fairness of the explanation $\boldsymbol{x}'$ as the expected distance between the counterfactual explanations of the original sample $\boldsymbol{x}_{orig}$ and such a perturbed sample $\boldsymbol{x}$:

$$\underset{\boldsymbol{x} \in \mathcal{B}_\epsilon(\boldsymbol{x}_{orig})}{\mathbb{E}} \left[\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \right] \tag{5}$$

Remark 1. Monte Carlo simulations, i.e. replacing the expectation in Eq. (5) with a sample mean gives us a method for empirically comparing the individual fairness (according to Definition 2) of different counterfactual explanations.

Note that the fairness criterion Eq. (5) is to be minimized - i.e. smaller values correspond to a better individual fairness. In contrast to the definition of individually fair causal recourse as proposed in [26], our definition(Definition 2) does not rely on the existence of a structural causal model, but it is formulated as a statistical criterion on the explanations (represented as samples).

3.2 Natural Perturbations

While there are different possible ways of perturbing a given input $\boldsymbol{x}$ - i.e. choosing $\mathcal{B}_\epsilon(\cdot)$ in Definition 2 -, we focus on three specific perturbations in this work: Perturbation by Gaussian noise, bounded uniform noise and a perturbation by masking features.

Gaussian Noise Perturbing a given input $\boldsymbol{x}$ with Gaussian noise means to add a small amount of normally distributed noise $\boldsymbol{\delta}$ to $\boldsymbol{x}$:

$$\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{\delta} \quad \text{where } \boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{6}$$

where the size and shape of the perturbation can be controlled by the covariance matrix $\mathbf{\Sigma}$ - in this work we use a diagonal matrix and often choose $\mathbf{\Sigma} = \mathbb{I}$. However, note that in this perturbation we can not guarantee that the perturbed sample is ϵ-close because $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ can yield arbitrarily large values. The probability of such values is limited, however.

Bounded Uniform Noise Similar to a perturbation with Gaussian noise (see Eq. (6)), we add a small amount of uniformly distributed noise $\boldsymbol{\delta}$ to $\boldsymbol{x}$:

$$\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{\delta} \quad \text{where } \boldsymbol{\delta} \sim \mathcal{U}(-\epsilon\mathbf{1}, \epsilon\mathbf{1}) \tag{7}$$

where $\epsilon > 0$ controls and upper bounds the amount of noise.

Note that, unlike a perturbation with Gaussian noise, we can guarantee that the perturbed sample is ϵ-close to the original sample.

Feature Masking While a perturbation by Gaussian noise Eq. (6) potentially changes every feature, feature masking allows a more precise way of perturbing a given input:

$$\boldsymbol{x} = \boldsymbol{x} \odot \boldsymbol{m} \quad \text{where } \boldsymbol{m} \in \{0,1\}^d \tag{8}$$

where $\odot$ denotes the element wise multiplication, and the size of the perturbation can be controlled by the number of masked features. Also note that the number of 0s (number of masked features) as well as their position (feature id) can vary - in this work, due to the absence of expert knowledge, given a fixed number of masked features, we select the masked features randomly.

3.3 Robustness of Counterfactual Explanations

In this section we formally study robustness and fairness of different prediction functions h. First, we give a general bound on the robustness of closest counterfactual explanations in Theorem 1 and then study each type of perturbation separately.

Theorem 1 (General bound on closest counterfactuals of perturbed samples). *Let $h : \mathcal{X} \to \mathcal{Y}$ be a prediction function and let $(\boldsymbol{x}_{orig}, y_{orig}) \in \mathcal{X} \times \mathcal{Y}$ be a sample for which we are given a closest counterfactual (see Eq. (2)) $\boldsymbol{x}' = CF(\boldsymbol{x}_{orig}, y') \in \mathcal{X}$. Let $\boldsymbol{x} \sim \mathcal{B}_\epsilon(\boldsymbol{x}_{orig})$ be a perturbed version of $\boldsymbol{x}_{orig}$ such that $\|\boldsymbol{x}_{orig} - \boldsymbol{x}\|_2 \leq \epsilon$ – we denote the corresponding closest counterfactual of $\boldsymbol{x}$ as $\boldsymbol{x}''$. We can then bound the difference between the two counterfactuals $\boldsymbol{x}'$ and $\boldsymbol{x}''$ as follows:*

$$\|\boldsymbol{x}' - \boldsymbol{x}''\|_2 \leq 2\epsilon + 2\|\boldsymbol{x}_{orig} - \boldsymbol{x}'\|_2 \tag{9}$$

In case of a binary linear classifier, we can refine the bound from Theorem 1 as stated in Corollary 1.

Corollary 1 (Bound on closest counterfactuals of perturbed samples for binary linear classifiers). *In case of a binary linear classifier, i.e. $h(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$ with $|\boldsymbol{w}| = 1$, we can refine the bound from Theorem 1 as follows:*

$$\|\boldsymbol{x}' - \boldsymbol{x}''\|_2 \leq 2\epsilon + 2|\boldsymbol{w}^\top \boldsymbol{x}_{orig}| \tag{10}$$

Remark 2. Note that while the general bound Eq. (9) in Theorem 1 depends on the closest counterfactual of the unperturbed sample $\boldsymbol{x}_{\text{orig}}$, the bound Eq. (10) in Corollary 1 only depends on the original sample $\boldsymbol{x}_{\text{orig}}$, the model parameter $\boldsymbol{w}$ and the perturbation bound ϵ. Both bounds (Theorem 1 and Corollary 1) are rather loose because they do not make any assumption on the perturbation $\mathcal{B}_\epsilon$ except that it must be bounded by ϵ - in addition the bound in Theorem 1 does not even make any assumption on h at all.

Gaussian Noise Additional assumptions allow more precise (and potentially more useful) statements as shown in the next theorem Theorem 2. In Theorem 2 (and the consequential corollaries Corollary 2 and Corollary 3) we study the individual fairness of a binary linear classifier under Gaussian noise. It turns out, that the individual fairness as defined in Definition 2 of closest counterfactual explanations of a binary linear classifier under Gaussian noise depends on the dimension d only – i.e. the larger the dimension of the input space, the larger the individual unfairness.

Theorem 2 (Individual fairness of closest counterfactuals of a linear binary classifier under Gaussian noise). *Let $h : \mathbb{R}^d \to \{-1, 1\}$ be a binary linear classifier – i.e. $h(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$ with $\|\boldsymbol{w}\|_2 = 1$. The individual fairness of closest counterfactuals (see Definition 2) under Gaussian noise Eq.* (6) *- with an arbitrary diagonal covariance $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_i^2)$ - at a sample $(\boldsymbol{x}_{orig}, y_{orig}) \in \mathbb{R}^d \times \{-1, 1\}$ can be stated as follows:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_{orig}, \boldsymbol{\Sigma})}\left[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\right] = \operatorname{trace}(\boldsymbol{\Sigma}) - \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \tag{11}$$

where we assume the squared Euclidean distance as a distance metric $\mathrm{d}(\cdot)$ for measuring the distance between two counterfactuals.

Corollary 2. *If we assume the identity matrix $\mathbb{I}$ as a covariance matrix $\boldsymbol{\Sigma}$ of the Gaussian noise in Theorem 2, Eq.* (11) *simplifies as follows:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_{orig}, \mathbb{I})}\left[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\right] = d - 1 \tag{12}$$

Corollary 3. *In the setting of Theorem 2 and Corollary 2, the probability that the individual unfairness is larger than a given $\delta > 0$ is bounded as follows:*

$$\mathbb{P}\Big(\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \geq \delta\Big) \leq \frac{d-1}{\delta} \tag{13}$$

Remark 3. We can interpret Theorem 2 (and in particular the consequential correlaries Corollary 2 and Corollary 3) as a "curse of dimensionality" for individual fairness of counterfactual explanations: the larger the dimension d of the data space, the larger the potential individual unfairness if the data manifold is intrinsically locally high dimensional.

Bounded Uniform Noise We can still make some statements on the individual fairness Definition 2 of closest counterfactual explanations, when using bounded uniform noise instead of Gaussian noise, as stated in Theorem 3 and Corollary 4.

Theorem 3 (Individual fairness of closest counterfactuals of a linear binary classifier under bounded uniform noise). *Let $h : \mathbb{R}^d \to \{-1, 1\}$ be a binary linear classifier - i.e. $h(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$ with $\|\boldsymbol{w}\|_2 = 1$. The individual fairness of closest counterfactuals (see Definition 2) under a bounded uniform noise Eq.* (7) *at an arbitrary sample $(\boldsymbol{x}_{orig}, y_{orig}) \in \mathbb{R}^d \times \{-1, 1\}$ can be stated as follows:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{U}(\boldsymbol{x}_{orig} \pm \epsilon \mathbf{1})}\left[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\right] = \frac{\epsilon^2 (d-1)}{3} \tag{14}$$

Corollary 4. *In the setting of Theorem 3, the probability that the individual unfairness is larger than some $\delta > 0$ can be upper bounded as:*

$$\mathbb{P}\Big(\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \geq \delta\Big) \leq \frac{\delta\epsilon^2(d-1)}{3} \tag{15}$$

Feature Masking In the previous theorems we studied general robustness and individual fairness Definition 2 of closest counterfactual explanations under Gaussian and bounded uniform noise. In Theorem 4 and Corollary 5 we study individual fairness under arbitrary feature masking (see Eq. (8)).

Theorem 4 (Individual fairness of closest counterfactuals of a linear binary classifier under arbitrary feature masking). *Let $h : \mathbb{R}^d \to \{-1, 1\}$ be a binary linear classifier - i.e. $h(\boldsymbol{x}) = \operatorname{sign}\big(\boldsymbol{w}^\top \boldsymbol{x}\big)$ with $\|\boldsymbol{w}\|_2 = 1$. The individual fairness of closest counterfactuals (see Definition 2) under arbitrary feature masking Eq. (8) at an arbitrary sample $(\boldsymbol{x}_{orig}, y_{orig}) \in \mathbb{R}^d \times \{-1, 1\}$ can be upper bounded as follows:*

$$\mathbb{E}\big[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\big] \leq (d+2)\|\boldsymbol{x}_{orig}\|_2^2 \tag{16}$$

Corollary 5. *In the setting of Theorem 4, the probability that the individual unfairness is larger than some $\delta > 0$ can be upper bounded as:*

$$\mathbb{P}\Big(\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \geq \delta\Big) \leq \frac{d+2}{\delta}\|\boldsymbol{x}_{orig}\|_2^2 \tag{17}$$

Remark 4. Note that the bounds in Theorem 4 and Corollary 5 are rather loose because we do not make any assumptions on the feature masking - neither on the number of masked feature nor on the distribution of the masked features. We could refine the bounds by making additional assumptions - see appendix A for details.

We will empirically confirm these mathematical findings of a potential presence of unfairness even for simple models in the experiments (see section 4). As a mediation, we propose to add a regularization for improving the individual fairness Definition 2 of counterfactual explanations. One solution can exist in the reference to plausible counterfactuals instead of closest ones. One reason of individual unfairness of closest counterfactuals comes from the fact that counterfactuals can populate all dimensions such that differences can accumulate in high dimensions. This can be avoided if data are regularized according to given observations. In addition, if a nonlinear rather than linear decision boundary is present, peculiarities of the decision boundary have less effect on the location of the counterfactuals this way.

4 Experiments

We empirically evaluate the individual fairness (Definition 2) of closest and plausible counterfactual explanations. Furthermore, we also empirically confirm the occurrence of the "curse of dimensionality" (see Remark 3) in case of non-linear classifiers. For this purpose, we compute closest and plausible counterfactuals of perturbed data points for a diverse set of classifiers and data sets.

Data sets We use an artificial toy data set and three standard data sets:

- Toy data set: A binary classification problem where each class is characterized by an isotropic Gaussian blob. The two blobs are slightly overlapping and the number of dimensions is varied - we use this data set for testing for the "curse of dimensionality" (see Remark 3).
- The "Breast Cancer Wisconsin (Diagnostic) Data Set" [45] whereby we add a PCA dimensionality reduction to 5 dimensions to the model.
- The "Wine data set" [37].
- The "Optical Recognition of Handwritten Digits Data Set" [2] whereby we add a PCA dimensionality reduction to 40 dimensions to the model.

Because PCA preprocessing is an affine transformation, we can integrate the transformation into the convex programs and therefore still compute counterfactuals in the original data space [7].

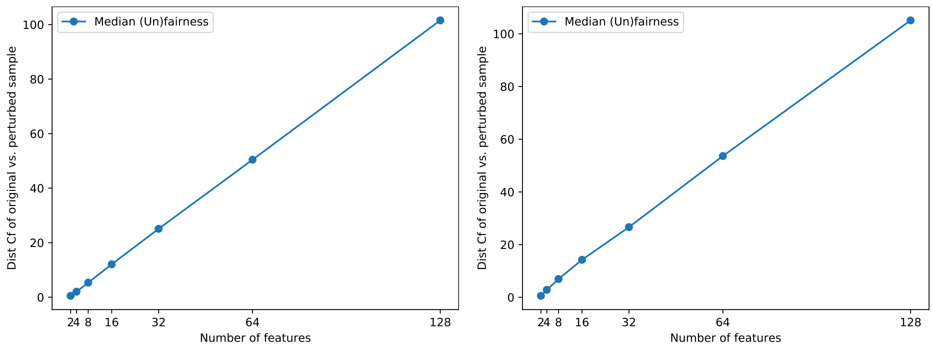
Models We use the following diverse set of models: softmax regression, generalized learning vector quantization (GLVQ) and decision tree classifier. We use the same hyperparameters across all data sets - for all vector quantization models we use 3 prototypes per class and use 7 as the maximum depth of decision tree classifiers.

Curse of Dimensionality In order to empirically study the occurrence of the "curse of dimensionality" (see Remark 3), we use the Gaussian blobs toy data set for fitting and evaluating counterfactuals (original vs. perturbed sample) under a decision tree classifier and a GLVQ model. The results (over a 4-fold cross validation) are shown in Fig. 1. Similar to the case of a linear classifier (see Theorem 2), we observe that even for non-linear classifiers, increasing the number of dimensions leads to an increase of the expected unfairness.

Setup - Closest vs. Plausible Counterfactuals We report the results of the following experiments over a 4-fold cross validation: We fit all models on the training data (depending on the data set this might involve a PCA as a preprocessing) and compute a closest and plausible counterfactual explanations of all samples from the test set that are classified correctly by the model - whereby we compute counterfactuals of the original as well as the perturbed sample. We use two different types of perturbations: Gaussian noise Eq. (6) with $\mathbf{\Sigma} = \mathbb{I}$ and feature masking Eq. (8) for one up to half of the total number of features. In case of a multi-class problem, we chose a random target label that is different from the original label. We compute and report the distance between the counterfactuals of the original sample and the perturbed sample Eq. (5) - we do this separately for closest and plausible counterfactuals. Furthermore, we use MOSEK[1] as a solver

[1] We gratefully acknowledge an academic license provided by MOSEK ApS.

Fig. 1: Gaussian blobs: Median absolute distance between counterfactual of original sample and perturbed sample (using Gaussian noise Eq. (6) with $\mathbf{\Sigma} = \mathbb{I}$). Smaller values are better.



Decision tree GLVQ

Data set	Wine		Breast cancer		Handwritten digits	
Method	Closest	Plausible	Closest	Plausible	Closest	Plausible
Softmax regression	10.16	**1.87**	24.04	**22.48**	53.71	**48.78**
Decision tree	9.25	**2.42**	24.05	**23.11**	56.56	**49.40**
GLVQ	9.95	**1.74**	23.34	**21.42**	57.66	**49.46**

Table 1: Comparing the median absolute distance between counterfactual of original sample and perturbed sample (using Gaussian noise Eq. (6)) - closest and plausible counterfactual explanations. Smaller values are better - best values are **highlighted**.

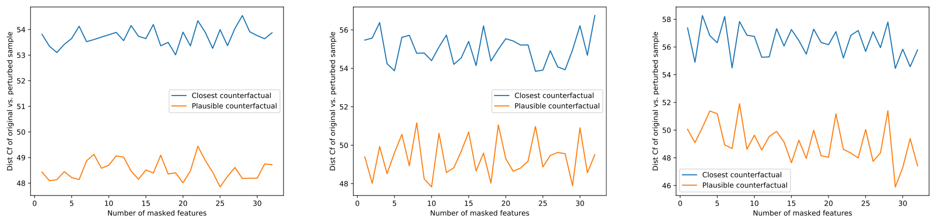
for all mathematical programs. The complete implementation of the experiments is available on GitHub[2].

Results The results of using Gaussian noise Eq. (6) for perturbing the samples are shown in Table 1. The results on the digit data set for increasingly masking more and more features Eq. (8) are shown in Fig. 2 - plots for the other data sets are given in appendix B.

We observe that in all cases the plausible counterfactual explanations are less affected by perturbations than the closest counterfactuals - thus we consider them to be better under individual fairness (Definition 2). The size of the differences depends a lot on the combination of model and data set. However, in all cases the difference is significant. In case of increasingly masking features, we observe that although the distance between counterfactuals of original and perturbed sample is subject to some variance, the difference between closest and plausible counterfactual is always significant - even when masking up to 50% of all features.

[2] https://github.com/andreArtelt/FairnessRobustnessContrastingExplanations

Fig. 2: Handwritten digits data set: Median absolute distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (8)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression | Decision tree | GLVQ

5 Discussion and Conclusion

In this work we argued that not only the fairness of decision making systems is important but also fairness of explanations is important. We studied the robustness of contrasting explanations - in particular we focused on counterfactual explanations. Besides deriving robustness bounds, we also focused on individual fairness of constrasting explanations - we studied formally and empirically the individual fairness of counterfactual explanations. In addition, we proposed to use plausible instead of closest counterfactuals for increasing the individual fairness of counterfactual explanations - we empirically evaluated and compared the individual fairness of closest vs. plausible counterfactual explanations. We found evidence that plausible counterfactuals provide better individual fairness than closest counterfactual explanations.

In future work we plan to further study formal fairness and robustness guarantees and bounds of more models (e.g. deep neural networks) and different perturbations. We also would like to investigate other approaches and methodologies for computing plausible counterfactual explanations - the work [8] we used in this work is only one possible approach for computing plausible counterfactuals, other approaches exist as well [29, 34]. Finally, we are highly interested in studying the problem of individual fairness of (contrasting) explanations from a psychological perspective - i.e. investigating how people actually experience individual fairness of contrasting explanations and whether this experience is successfully captured/modelled by our proposed formalizations and methods.

References

1. Aamodt, A., Plaza., E.: Case-based reasoning: Foundational issues, methodological variations, and systemapproaches. AI communications (1994)
2. Alpaydin, E., Kaynak, C.: Optical recognition of handwritten digits data set. https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits (1998)

3. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. CoRR **abs/1806.08049** (2018), http://arxiv.org/abs/1806.08049
4. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 7786–7795 (2018), https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html
5. Anders, C.J., Pasliev, P., Dombrowski, A., Müller, K., Kessel, P.: Fairwashing explanations with off-manifold detergent. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 314–323. PMLR (2020), http://proceedings.mlr.press/v119/anders20a.html
6. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias - theres software used across the country to predict future criminals. and its biased against blacks. (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
7. Artelt, A., Hammer, B.: On the computation of counterfactual explanations - A survey. CoRR **abs/1911.07749** (2019), http://arxiv.org/abs/1911.07749
8. Artelt, A., Hammer, B.: Convex density constraints for computing plausible counterfactual explanations. 29th International Conference on Artificial Neural Networks (ICANN) (2020)
9. Binns, R.: On the apparent conflict between individual and group fairness. In: Hildebrandt, M., Castillo, C., Celis, E., Ruggieri, S., Taylor, L., Zanfir-Fortuna, G. (eds.) FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020. pp. 514–524. ACM (2020). https://doi.org/10.1145/3351095.3372864, https://doi.org/10.1145/3351095.3372864
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, NY, USA (2004)
11. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 6276–6282. International Joint Conferences on Artificial Intelligence Organization (7 2019). https://doi.org/10.24963/ijcai.2019/876, https://doi.org/10.24963/ijcai.2019/876
12. Caton, S., Haas, C.: Fairness in machine learning: A survey. CoRR **abs/2010.04053** (2020), https://arxiv.org/abs/2010.04053
13. Chang, H., Nguyen, T.D., Murakonda, S.K., Kazemi, E., Shokri, R.: On adversarial bias and the robustness of fair machine learning. CoRR **abs/2006.08669** (2020), https://arxiv.org/abs/2006.08669
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Goldwasser, S. (ed.) Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012. pp. 214–226. ACM (2012). https://doi.org/10.1145/2090236.2090255, https://doi.org/10.1145/2090236.2090255
15. Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. arXiv e-prints arXiv:1801.01489 (Jan 2018)

16. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018. pp. 80–89 (2018). https://doi.org/10.1109/DSAA.2018.00018, https://doi.org/10.1109/DSAA.2018.00018
17. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. CoRR **abs/1805.04755** (2018), http://arxiv.org/abs/1805.04755
18. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. CoRR **abs/1805.10820** (2018), http://arxiv.org/abs/1805.10820
19. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (Aug 2018). https://doi.org/10.1145/3236009, http://doi.acm.org/10.1145/3236009
20. Gupta, V., Nokhiz, P., Roy, C.D., Venkatasubramanian, S.: Equalizing recourse across groups. CoRR **abs/1909.03166** (2019), http://arxiv.org/abs/1909.03166
21. Hancox-Li, L.: Robustness in machine learning explanations: Does it matter? p. 640647. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372836, https://doi.org/10.1145/3351095.3372836
22. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 2921–2932 (2019), https://proceedings.neurips.cc/paper/2019/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html
23. Kim, B., Koyejo, O., Khanna, R.: Examples are not enough, learn to criticize! criticism for interpretability. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 2280–2288 (2016)
24. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. pp. 1885–1894 (2017), http://proceedings.mlr.press/v70/koh17a.html
25. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4066–4076 (2017), https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
26. von Kgelgen, J., Karimi, A.H., Bhatt, U., Valera, I., Weller, A., Schlkopf, B.: On the fairness of causal algorithmic recourse (2021)
27. Laugel, T., Lesot, M., Marsala, C., Detyniecki, M.: Issues with post-hoc counterfactual explanations: a discussion. CoRR **abs/1906.04774** (2019), http://arxiv.org/abs/1906.04774
28. Laugel, T., Renard, X., Lesot, M., Marsala, C., Detyniecki, M.: Defining locality for surrogates in post-hoc interpretablity. CoRR **abs/1806.07498** (2018), http://arxiv.org/abs/1806.07498

29. Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. CoRR **abs/1907.02584** (2019)
30. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. CoRR **abs/1908.09635** (2019), http://arxiv.org/abs/1908.09635
31. Molnar, C.: Interpretable Machine Learning (2019), https://christophm.github.io/interpretable-ml-book/
32. Nanda, V., Dooley, S., Singla, S., Feizi, S., Dicker-son., J.P.: Fairness through robustness:investigating robustness disparity in deep learning. In: FAccT'21: Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, March 310, 2021. ACM (2021). https://doi.org/10.1145/3442188.3445910, https://doi.org/10.1145/3442188.3445910
33. parliament, E., council: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016)
34. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., Bie, T.D., Flach, P.A.: FACE: feasible and actionable counterfactual explanations. CoRR **abs/1909.09369** (2019), http://arxiv.org/abs/1909.09369
35. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. In: ICML Workshop on Human Interpretability in Machine Learning (WHI) (2016)
36. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. KDD '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, http://doi.acm.org/10.1145/2939672.2939778
37. S. Aeberhard, D.C., de Vel, O.: Comparison of classifiers in high dimensional settings. Tech. Rep. no. 92-02 (1992)
38. Samek, W., Wiegand, T., Müller, K.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. CoRR **abs/1708.08296** (2017), http://arxiv.org/abs/1708.08296
39. Sokol, K., Flach, P.A.: Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In: Espinoza, H., hÉigeartaigh, S.Ó., Huang, X., Hernández-Orallo, J., Castillo-Effen, M. (eds.) Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019. CEUR Workshop Proceedings, vol. 2301. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2301/paper_20.pdf
40. Stevens, A., Deruyck, P., Veldhoven, Z.V., Vanthienen, J.: Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In: 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, Canberra, Australia, December 1-4, 2020. pp. 1241–1248. IEEE (2020). https://doi.org/10.1109/SSCI47803.2020.9308371, https://doi.org/10.1109/SSCI47803.2020.9308371
41. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): towards medical XAI. CoRR **abs/1907.07374** (2019), http://arxiv.org/abs/1907.07374
42. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review (2020)

43. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR **abs/1711.00399** (2017), http://arxiv.org/abs/1711.00399
44. Waddell, K.: How algorithms can bring down minorities' credit scores. The Atlantic (2016)
45. William H. Wolberg, W. Nick Street, O.L.M.: Breast cancer wisconsin (diagnostic) data set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (1995)
46. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. Journal of Business & Economic Statistics **0**(ja), 1–19 (2019). https://doi.org/10.1080/07350015.2019.1624293, https://doi.org/10.1080/07350015.2019.1624293

A Proofs and Derivations

1. *Proof (Theorem 1).* Since the perturbation is bounded by $\epsilon > 0$, it holds that:

$$\|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}\|_2 \leq \epsilon \tag{18}$$

Furthermore, if the closest counterfactual $\boldsymbol{x}'$ of $\boldsymbol{x}_{\text{orig}}$ is different from the closest counterfactual $\boldsymbol{x}''$ of $\boldsymbol{x}$ (perturbed $\boldsymbol{x}_{\text{orig}}$), it must hold that:

$$\|\boldsymbol{x} - \boldsymbol{x}''\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \tag{19}$$

Because of the triangle inequality we know that the following holds:

$$\|\boldsymbol{x}'' - \boldsymbol{x}'\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}''\|_2 + \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \tag{20}$$

Plugging Eq. (19) into Eq. (20) yields:

$$\begin{aligned} \|\boldsymbol{x}' - \boldsymbol{x}''\|_2 &\leq \|\boldsymbol{x} - \boldsymbol{x}''\|_2 + \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{x}'\|_2 + \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \\ &= 2\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \end{aligned} \tag{21}$$

By making use of the triangle inequality and Eq. (18), we find that:

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 &\leq \|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}\|_2 + \|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}'\|_2 \\ &\leq \epsilon + \|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}'\|_2 \end{aligned} \tag{22}$$

Plugging Eq. (22) into Eq. (21) yields the desired bound Eq. (9):

$$\begin{aligned} \|\boldsymbol{x}' - \boldsymbol{x}''\|_2 &\leq 2\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \\ &= 2\epsilon + 2\|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}'\|_2 \end{aligned} \tag{23}$$

□

2. *Proof (Corollary 1).* First, we prove that the closest counterfactual explanations $\boldsymbol{x}'$ of a sample $\boldsymbol{x}_{\text{orig}}$ under a binary linear classifier $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x})$ (we assume w.l.o.g. $\|\boldsymbol{w}\|_2 = 1$) can be explicitly stated as follows:

$$\boldsymbol{x}' = \boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right) \boldsymbol{w} \tag{24}$$

Computing the closest counterfactual of some $\boldsymbol{x}_{\text{orig}}$ under a binary linear classifier can be formalized as the following optimization problem:

$$\min_{\boldsymbol{x}' \in \mathbb{R}^d} \|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}'\|_2^2 \tag{25a}$$

$$\text{s.t. } \boldsymbol{w}^\top \boldsymbol{x}' = 0 \tag{25b}$$

Note that the constraint Eq. (25b) "replaces/approximates" the constraint $h(\boldsymbol{x}') = y'$ Eq. (2b). The constraint Eq. (25b) requires that the solution $\boldsymbol{x}'$ lies directly on the decision boundary. We assume that points on the decision boundary are classified as y' - while this approach is debatable, it offers an easy solution to the original problem because otherwise we would have to project onto an open set which is "difficult" (once we are on the decision boundary we could add an infinitesimally small constant to the solution for crossing the decision boundary if this is really necessary).

We solve Eq. (25) by using the method of Lagrangian multipliers. Since Eq. (25) is a convex optimization problem, we only have globally optimal solutions - in particular we have a quadratic objective and an linear constraint, thus strong duality holds. The Lagrangian of Eq. (25) is given as follows:

$$\mathcal{L}(\boldsymbol{x}', \lambda) = \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}' + \boldsymbol{x}'^\top \boldsymbol{x}' - \lambda \boldsymbol{w}^\top \boldsymbol{x}' \tag{26}$$

The gradient of the Lagrangian Eq. (26) with respect to $\boldsymbol{x}'$ can be written as follows:

$$\begin{aligned} \nabla_{\boldsymbol{x}'} \mathcal{L}(\boldsymbol{x}', \lambda) &= \nabla_{\boldsymbol{x}'} \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - \nabla_{\boldsymbol{x}'} 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}' + \nabla_{\boldsymbol{x}'} \boldsymbol{x}'^\top \boldsymbol{x}' - \nabla_{\boldsymbol{x}'} \lambda \boldsymbol{w}^\top \boldsymbol{x}' \\ &= -2\boldsymbol{x}_{\text{orig}} + 2\boldsymbol{x}' - \lambda \boldsymbol{w} \end{aligned} \tag{27}$$

The optimality condition requires the gradient Eq. (27) being equal to zero:

$$\begin{aligned} &\nabla_{\boldsymbol{x}'} \mathcal{L}(\boldsymbol{x}', \lambda) = \boldsymbol{0} \\ &\Leftrightarrow -2\boldsymbol{x}_{\text{orig}} + 2\boldsymbol{x}' - \lambda \boldsymbol{w} = 0 \\ &\Leftrightarrow \boldsymbol{x}' = \boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2} \boldsymbol{w} \end{aligned} \tag{28}$$

Plugging Eq. (28) back into the Lagrangian Eq. (26) yields the Lagrangian dual:

$$
\begin{aligned}
\mathcal{L}_D(\lambda) &= \min_{\boldsymbol{x}' \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{x}', \lambda) \\
\mathcal{L}\left(\boldsymbol{x}' = \boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w}, \lambda\right) &= \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \left(\boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w}\right) + \\
&\quad \left(\boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w}\right)^\top \left(\boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w}\right) - \lambda \boldsymbol{w}^\top \left(\boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w}\right) \\
&= \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - \lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} + \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} + \\
&\quad \lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} + \frac{\lambda^2}{4}\boldsymbol{w}^\top \boldsymbol{w} - \lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} - \frac{\lambda^2}{2}\boldsymbol{w}^\top \boldsymbol{w} \\
&= \frac{\lambda^2}{4}\boldsymbol{w}^\top \boldsymbol{w} - \lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} - \frac{\lambda^2}{2}\boldsymbol{w}^\top \boldsymbol{w} \\
&= -\frac{\lambda^2}{4}\boldsymbol{w}^\top \boldsymbol{w} - \lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}
\end{aligned}
\tag{29}
$$

The gradient of the Lagrangian dual Eq. (29) can be written as follows:

$$
\begin{aligned}
\frac{\partial}{\partial \lambda}\mathcal{L}_D(\lambda) &= -\frac{\partial}{\partial \lambda}\frac{\lambda^2}{4}\boldsymbol{w}^\top \boldsymbol{w} - \frac{\partial}{\partial \lambda}\lambda \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} \\
&= -\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} - \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}
\end{aligned}
\tag{30}
$$

Next, the optimality condition requires that the gradient Eq. (30) is equal to zero:

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda}\mathcal{L}_D(\lambda) = 0 \\
&\Leftrightarrow -\frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} - \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} = 0 \\
&\Leftrightarrow \lambda = -2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w} = -2\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}
\end{aligned}
\tag{31}
$$

where we made use of $\|\boldsymbol{w}\|_2 = 1$.
Finally, we obtain the solution of the original problem Eq. (25) by plugging the solution of the dual problem Eq. (31) into Eq. (28):

$$
\begin{aligned}
\boldsymbol{x}' &= \boldsymbol{x}_{\text{orig}} + \frac{\lambda}{2}\boldsymbol{w} \\
&= \boldsymbol{x}_{\text{orig}} + \frac{-2\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}}{2}\boldsymbol{w} \\
&= \boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right)\boldsymbol{w}
\end{aligned}
\tag{32}
$$

which concludes this sub-proof.

Plugging Eq. (24) into the bound Eq. (9) from Theorem 1, and again assuming w.l.o.g. that $\|\boldsymbol{w}\|_2 = 1$, yields the desired bound Eq. (10):

$$\begin{aligned}
\|\boldsymbol{x}' - \boldsymbol{x}''\|_2 &\leq 2\epsilon + 2\|\boldsymbol{x}_{\text{orig}} - \boldsymbol{x}'\|_2 \\
&= 2\epsilon + 2\|\boldsymbol{x}_{\text{orig}} - (\boldsymbol{x}_{\text{orig}} - \boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\boldsymbol{w})\|_2 \\
&= 2\epsilon + 2\|\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\boldsymbol{w}\|_2 \\
&= 2\epsilon + 2|\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}|
\end{aligned} \tag{33}$$

□

3. *Proof (Theorem 2).* From the proof of Corollary 1 we now that the closest counterfactual explanation $\boldsymbol{x}'$ of a sample $\boldsymbol{x}$ under a linear binary classifier $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ can be stated explicitly Eq. (24):

$$\boldsymbol{x}' = \boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}\right) \boldsymbol{w} \tag{34}$$

Applying the analytic solution Eq. (34) to the squared Euclidean distance between the closest counterfactual $\boldsymbol{x}'$ of the original sample $\boldsymbol{x}_{\text{orig}}$ and the closest counterfactual $\boldsymbol{x}''$ of the corresponding perturbed sample $\boldsymbol{x}$ yields:

$$\begin{aligned}
\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') &= \left(\boldsymbol{x}' - \boldsymbol{x}''\right)^\top \left(\boldsymbol{x}' - \boldsymbol{x}''\right) \\
&= \boldsymbol{x}'^\top \boldsymbol{x}' - 2\boldsymbol{x}'^\top \boldsymbol{x}'' + \boldsymbol{x}''^\top \boldsymbol{x}'' \\
&= \left(\boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right) \boldsymbol{w}\right)^\top \left(\boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right) \boldsymbol{w}\right) - \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right) \boldsymbol{w}\right)^\top \left(\boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}\right) \boldsymbol{w}\right) + \\
&\quad \left(\boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}\right) \boldsymbol{w}\right)^\top \left(\boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}\right) \boldsymbol{w}\right) \\
&= \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x} - 2\left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right)^2 + 2\boldsymbol{x}_{\text{orig}}^\top \left(\boldsymbol{w}^\top \boldsymbol{x}\right) \boldsymbol{w} + \\
&\quad \boldsymbol{x}^\top \boldsymbol{x} + 2\left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right)\left(\boldsymbol{x}^\top \boldsymbol{w}\right) - 2\left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2 + \left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right)^2 - \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right)^\top \left(\boldsymbol{x}^\top \boldsymbol{w}\right) + \left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2 \\
&= \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right)^2 + 2\left(\boldsymbol{w}^\top \boldsymbol{x}\right)\left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right) + \\
&\quad \boldsymbol{x}^\top \boldsymbol{x} - \left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2
\end{aligned} \tag{35}$$

Taking the expectation of Eq. (35) over an arbitrary density p yields:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p}\left[\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\right] &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p}\Big[\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{w}^\top \boldsymbol{x}\right)\left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right) + \boldsymbol{x}^\top \boldsymbol{x} - \left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2\Big] \\
&= \boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}_{\text{orig}} - \mathbb{E}\Big[2\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{x}\Big] - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}}^\top \boldsymbol{w}\right)\mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}\Big] + \mathbb{E}\Big[\boldsymbol{x}^\top \boldsymbol{x}\Big] - \mathbb{E}\Big[\left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2\Big]
\end{aligned} \quad y \tag{36}$$

Working out the specific expectations from Eq. (36) and under a Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_{\text{orig}}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ - i.e. $(\boldsymbol{x})_i$s are uncorrelated

- yields:

$$\begin{aligned}\mathbb{E}\Big[2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[2\sum_i(\boldsymbol{x}_{\text{orig}})_i(\boldsymbol{x})_i\Big] \\ &= 2\sum_i(\boldsymbol{x}_{\text{orig}})_i\mathbb{E}\big[(\boldsymbol{x})_i\big] \\ &= 2\sum_i(\boldsymbol{x}_{\text{orig}})_i(\boldsymbol{x}_{\text{orig}})_i \\ &= 2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}}\end{aligned} \tag{37}$$

$$\begin{aligned}\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[\sum_i(\boldsymbol{w})_i(\boldsymbol{x})_i\Big] \\ &= \sum_i(\boldsymbol{w})_i\mathbb{E}\big[(\boldsymbol{x})_i\big] \\ &= \sum_i(\boldsymbol{w})_i(\boldsymbol{x}_{\text{orig}})_i \\ &= \boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\end{aligned} \tag{38}$$

$$\begin{aligned}\mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[\sum_i(\boldsymbol{x})_i^2\Big] \\ &= \sum_i\mathbb{E}\big[(\boldsymbol{x})_i^2\big] \\ &= \sum_i\Big(\mathbb{E}\big[(\boldsymbol{x})_i\big]^2 + \text{Var}\,\big[(\boldsymbol{x})_i\big]\Big) \\ &= \sum_i\Big((\boldsymbol{x}_{\text{orig}})_i^2 + \sigma_i^2\Big) \\ &= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} + \text{trace}(\boldsymbol{\Sigma})\end{aligned} \tag{39}$$

$$\begin{aligned}\mathbb{E}\Big[\big(\boldsymbol{x}^\top\boldsymbol{w}\big)^2\Big] &= \mathbb{E}\big[\boldsymbol{x}^\top\boldsymbol{w}\big]^2 + \text{Var}\,\big[\boldsymbol{x}^\top\boldsymbol{w}\big] \\ &= \big(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\big)^2 + \text{Var}\,\Big[\sum_i(\boldsymbol{w})_i(\boldsymbol{x})_i\Big] \\ &= \big(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\big)^2 + \sum_i\text{Var}\,\big[(\boldsymbol{w})_i(\boldsymbol{x})_i\big] \\ &= \big(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\big)^2 + \sum_i(\boldsymbol{w})_i^2\sigma_i^2 \\ &= \big(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\big)^2 + \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\end{aligned} \tag{40}$$

where we made use of the assumption that $\|\boldsymbol{w}\|_2 = 1$.

Substituting Eq. (37), Eq. (38), Eq. (39), Eq. (40) in Eq. (36) yields:

$$\begin{aligned}
\underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}_{\text{orig}},\boldsymbol{\Sigma})}{\mathbb{E}}\big[\,\mathrm{d}(\boldsymbol{x}',\boldsymbol{x}'')\big] &= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} - \mathbb{E}\Big[2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}\Big] - \left(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{w}\right)\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}\Big] + \mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{x}\Big] - \mathbb{E}\Big[\left(\boldsymbol{x}^\top\boldsymbol{w}\right)^2\Big] \\
&= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} - 2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} - \left(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{w}\right)^2 + \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} + \operatorname{trace}(\boldsymbol{\Sigma}) - \left(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\right)^2 - \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w} \\
&= \operatorname{trace}(\boldsymbol{\Sigma}) - \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}
\end{aligned} \tag{41}$$

which concludes the proof. □

4. *Proof (Corollary 2).* Substituting $\mathbb{I}$ for $\boldsymbol{\Sigma}$ in Eq. (11) from Theorem 2 yields the claimed expectation:

$$\begin{aligned}
\underset{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}_{\text{orig}},\boldsymbol{\Sigma})}{\mathbb{E}}\big[\,\mathrm{d}(\boldsymbol{x}',\boldsymbol{x}'')\big] &= \operatorname{trace}(\boldsymbol{\Sigma}) - \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w} \\
&= \operatorname{trace}(\mathbb{I}) - \boldsymbol{w}^\top\mathbb{I}\boldsymbol{w} \\
&= d - 1
\end{aligned} \tag{42}$$

□

5. *Proof (Corollary 3).* Plugging the expectation from Corollary 2 into Markov's inequality yields the claimed bound:

$$\begin{aligned}
\mathbb{P}\Big(\,\mathrm{d}(\boldsymbol{x}',\boldsymbol{x}'') \geq \delta\Big) &\leq \frac{\mathbb{E}\big[\,\mathrm{d}(\boldsymbol{x}',\boldsymbol{x}'')\big]}{\delta} \\
&= \frac{d-1}{\delta}
\end{aligned} \tag{43}$$

□

6. *Proof (Theorem 3).* From the proof of Theorem 2 we know that the exceptation over an arbitrary density p of the distance between the closest counterfactual of the original sample and the perturbed sample can be written as follows:

$$\begin{aligned}
\underset{\boldsymbol{x}\sim p}{\mathbb{E}}\big[\,\mathrm{d}(\boldsymbol{x}',\boldsymbol{x}'')\big] &= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} - \mathbb{E}\Big[2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}\Big] - \left(\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{w}\right)\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}\Big] + \mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{x}\Big] - \mathbb{E}\Big[\left(\boldsymbol{x}^\top\boldsymbol{w}\right)^2\Big]
\end{aligned} \tag{44}$$

Next, working out the specific expectations from Eq. (44) under a bounded uniform noise $\boldsymbol{x} \sim \mathcal{U}(\boldsymbol{x}_{\text{orig}} \pm \epsilon\mathbf{1})$ - i.e. $(\boldsymbol{x})_i$s are uncorrelated - yields:

$$
\begin{aligned}
\mathbb{E}\Big[2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[2\sum_i(\boldsymbol{x}_{\text{orig}})_i(\boldsymbol{x})_i\Big] \\
&= 2\sum_i(\boldsymbol{x}_{\text{orig}})_i\mathbb{E}\big[(\boldsymbol{x})_i\big] \\
&= 2\sum_i(\boldsymbol{x}_{\text{orig}})_i\frac{1}{2}\Big((\boldsymbol{x}_{\text{orig}})_i - \epsilon + (\boldsymbol{x}_{\text{orig}})_i + \epsilon\Big) \\
&= 2\sum_i(\boldsymbol{x}_{\text{orig}})_i\Big(\boldsymbol{x}_{\text{orig}})_i \\
&= 2\boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}}
\end{aligned}
\tag{45}
$$

$$
\begin{aligned}
\mathbb{E}\Big[\boldsymbol{w}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[\sum_i(\boldsymbol{w})_i(\boldsymbol{x})_i\Big] \\
&= \sum_i(\boldsymbol{w})_i\mathbb{E}\big[(\boldsymbol{x})_i\big] \\
&= \sum_i(\boldsymbol{w})_i(\boldsymbol{x}_{\text{orig}})_i \\
&= \boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}}
\end{aligned}
\tag{46}
$$

$$
\begin{aligned}
\mathbb{E}\Big[\boldsymbol{x}^\top\boldsymbol{x}\Big] &= \mathbb{E}\Big[\sum_i(\boldsymbol{x})_i^2\Big] \\
&= \sum_i\mathbb{E}\big[(\boldsymbol{x})_i^2\big] \\
&= \sum_i\Big(\mathbb{E}\big[(\boldsymbol{x})_i\big]^2 + \mathrm{Var}\left[(\boldsymbol{x})_i\right]\Big) \\
&= \sum_i\Big((\boldsymbol{x}_{\text{orig}})_i^2 + \frac{1}{12}\Big((\boldsymbol{x}_{\text{orig}})_i + \epsilon - (\boldsymbol{x}_{\text{orig}})_i + \epsilon\Big)^2\Big) \\
&= \sum_i\Big((\boldsymbol{x}_{\text{orig}})_i^2 + \frac{4\epsilon^2}{12}\Big) \\
&= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} + \frac{d\epsilon^2}{3}
\end{aligned}
\tag{47}
$$

$$
\begin{aligned}
\mathbb{E}\Big[\left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2\Big] &= \mathbb{E}\big[\boldsymbol{x}^\top \boldsymbol{w}\big]^2 + \mathrm{Var}\left[\boldsymbol{x}^\top \boldsymbol{w}\right] \\
&= \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \mathrm{Var}\left[\sum_i (\boldsymbol{w})_i (\boldsymbol{x})_i\right] \\
&= \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \sum_i \mathrm{Var}\left[(\boldsymbol{w})_i (\boldsymbol{x})_i\right] \\
&= \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \sum_i (\boldsymbol{w})_i^2 \frac{\epsilon^2}{3} \\
&= \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \boldsymbol{w}^\top \boldsymbol{w} \frac{\epsilon^2}{3} \\
&= \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \frac{\epsilon^2}{3}
\end{aligned} \tag{48}
$$

Substituting Eq. (45), Eq. (46), Eq. (47), Eq. (48) in Eq. (44) yields:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{U}(\boldsymbol{x}_{\mathrm{orig}} \pm \epsilon \boldsymbol{1})}\big[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\big] &= \boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{x}_{\mathrm{orig}} - \mathbb{E}\Big[2\boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{x}\Big] - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{w}\right) \mathbb{E}\Big[\boldsymbol{w}^\top \boldsymbol{x}\Big] + \mathbb{E}\Big[\boldsymbol{x}^\top \boldsymbol{x}\Big] - \mathbb{E}\Big[\left(\boldsymbol{x}^\top \boldsymbol{w}\right)^2\Big] \\
&= \boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{x}_{\mathrm{orig}} - 2\boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{x}_{\mathrm{orig}} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 + \\
&\quad 2\left(\boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{w}\right)^2 + \boldsymbol{x}_{\mathrm{orig}}^\top \boldsymbol{x}_{\mathrm{orig}} + \frac{d\epsilon^2}{3} - \left(\boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{orig}}\right)^2 - \frac{\epsilon^2}{3} \\
&= \frac{\epsilon^2(d-1)}{3}
\end{aligned} \tag{49}
$$

which concludes the proof. □

7. *Proof (Corollary 4).* Plugging the expectation from Theorem 3 into Markov's inequality yields the claimed bound:

$$
\begin{aligned}
\mathbb{P}\Big(\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \geq \delta\Big) &\leq \frac{\mathbb{E}\big[\,\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\big]}{\delta} \\
&= \frac{\delta\epsilon^2(d-1)}{3}
\end{aligned} \tag{50}
$$

□

8. *Proof (Theorem 4).* Using the squared Euclidean distance and applying the triangle inequality yields:

$$
\begin{aligned}
\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') &= \|\boldsymbol{x}' - \boldsymbol{x}''\|_2^2 \\
&\leq \|\boldsymbol{x}'\|_2^2 + \|-\boldsymbol{x}''\|_2^2
\end{aligned} \tag{51}
$$

Applying the analytic solution of the closest counterfactual Eq. (24) and the triangle inequality to the second term in Eq. (51) yields:

$$
\begin{aligned}
\|-\boldsymbol{x}''\|_2^2 &= \|-\boldsymbol{x}_{\mathrm{orig}} \odot \boldsymbol{m} + \left(\boldsymbol{w}^\top (\boldsymbol{x}_{\mathrm{orig}} \odot \boldsymbol{m})\right) \boldsymbol{w}\|_2^2 \\
&\leq \|\boldsymbol{x}_{\mathrm{orig}} \odot \boldsymbol{m}\|_2^2 + \|\left(\boldsymbol{w}^\top (\boldsymbol{x}_{\mathrm{orig}} \odot \boldsymbol{m})\right) \boldsymbol{w}\|_2^2
\end{aligned} \tag{52}
$$

Applying the Cauchy-Schwarz inequality several times to Eq. (52) yields:

$$
\begin{aligned}
\|\left(\boldsymbol{w}^\top(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})\right)\boldsymbol{w}\|_2^2 &= \left(\sum_i (\boldsymbol{w})_i \boldsymbol{w}^\top(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})\right)^2 \\
&\leq \left(\sum_i (\boldsymbol{w})_i^2\right)\left(\sum_i (\boldsymbol{w}^\top(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}))^2\right) \\
&= d\left(\boldsymbol{w}^\top(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})\right)^2 \\
&= d\left(\sum_i (\boldsymbol{w})_i(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})_i\right)^2 \\
&\leq d\left(\sum_i (\boldsymbol{w})_i^2\right)\left(\sum_i (\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})_i^2\right) \\
&= d\left(\sum_i (\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})_i^2\right) \\
&= d\|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2
\end{aligned}
\tag{53}
$$

Substituting Eq. (53) back into Eq. (52) yields:

$$
\begin{aligned}
\|-\boldsymbol{x}''\|_2^2 &\leq \|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2 + \|\left(\boldsymbol{w}^\top(\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m})\right)\boldsymbol{w}\|_2^2 \\
&\leq \|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2 + d\|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2 \\
&= (d+1)\|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2
\end{aligned}
\tag{54}
$$

The second term in Eq. (54) can be upper bounded as follows:

$$
\begin{aligned}
\|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2 &= \sum_{j:\,(\boldsymbol{m})_j=1} (\boldsymbol{x}_{\text{orig}})_j^2 \\
&\leq \|\boldsymbol{x}_{\text{orig}}\|_2^2 - \sum_{j\in\operatorname{argsort}\left((\boldsymbol{x}_{\text{orig}})_i^2\right)[:d-\|\boldsymbol{m}\|_1]} (\boldsymbol{x}_{\text{orig}})_j^2
\end{aligned}
\tag{55}
$$

Note that this bound could be tightened by making additional assumptions on the feature mask $\boldsymbol{m}$ - e.g. number of masked features.
By applying the squared Euclidean norm to the closest counterfactual of the original sample $\boldsymbol{x}_{\text{orig}}$ as given in Eq. (24) yields:

$$
\begin{aligned}
\|\boldsymbol{x}'\|_2^2 &= \|\boldsymbol{x}_{\text{orig}} - (\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}})\boldsymbol{w}\|_2^2 \\
&= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} + (\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}})^2\boldsymbol{w}^\top\boldsymbol{w} - 2\boldsymbol{x}_{\text{orig}}^\top\left((\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}})\boldsymbol{w}\right) \\
&= \boldsymbol{x}_{\text{orig}}^\top\boldsymbol{x}_{\text{orig}} - (\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}})^2 \\
&= \|\boldsymbol{x}_{\text{orig}}\|_2^2 - (\boldsymbol{w}^\top\boldsymbol{x}_{\text{orig}})^2
\end{aligned}
\tag{56}
$$

Substituting everything back into Eq. (51) yields the following upper bound:

$$
\begin{aligned}
\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') &\leq \|\boldsymbol{x}'\|_2^2 + \|-\boldsymbol{x}''\|_2^2 \\
&= \|\boldsymbol{x}_{\text{orig}}\|_2^2 - (\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}})^2 + \|-\boldsymbol{x}''\|_2^2 \\
&\leq \|\boldsymbol{x}_{\text{orig}}\|_2^2 - (\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}})^2 + (d+1)\|\boldsymbol{x}_{\text{orig}} \odot \boldsymbol{m}\|_2^2 \\
&\leq \|\boldsymbol{x}_{\text{orig}}\|_2^2 - (\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}})^2 + (d+1)\left(\|\boldsymbol{x}_{\text{orig}}\|_2^2 - \sum_{j \in \operatorname{argsort}\left((\boldsymbol{x}_{\text{orig}})_i^2\right)[:d-\|\boldsymbol{m}\|_1]} (\boldsymbol{x}_{\text{orig}})_j^2\right) \\
&= (d+2)\|\boldsymbol{x}_{\text{orig}}\|_2^2 - (\boldsymbol{w}^\top \boldsymbol{x}_{\text{orig}})^2 - (d+1) \sum_{j \in \operatorname{argsort}\left((\boldsymbol{x}_{\text{orig}})_i^2\right)[:d-\|\boldsymbol{m}\|_1]} (\boldsymbol{x}_{\text{orig}})_j^2 \\
&\leq (d+2)\|\boldsymbol{x}_{\text{orig}}\|_2^2
\end{aligned}
\tag{57}
$$

Again, note that we could tighten this bound by making additional assumptions on the feature mask $\boldsymbol{m}$ as shown in Eq. (55).
Finally, the upper bound of the expectation of Eq. (57) over an arbitrary feature mask $\boldsymbol{m}$ follows immediately:

$$
\mathbb{E}\big[\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\big] \leq (d+2)\|\boldsymbol{x}_{\text{orig}}\|_2^2 \tag{58}
$$

which concludes the proof. □

9. *Proof (Corollary 5).* Plugging the upper bound of the expectation from Theorem 4 into Markov's inequality yields the claimed bound:

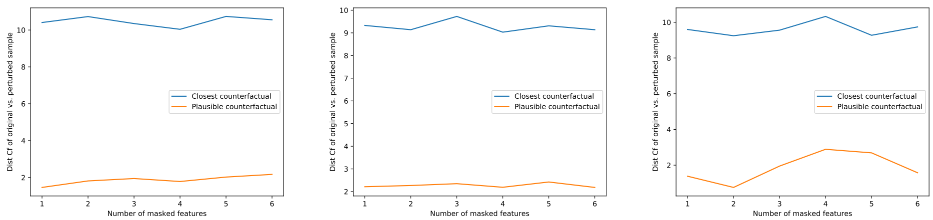
$$
\begin{aligned}
\mathbb{P}\Big(\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'') \geq \delta\Big) &\leq \frac{\mathbb{E}\big[\mathrm{d}(\boldsymbol{x}', \boldsymbol{x}'')\big]}{\delta} \\
&= \frac{d+2}{\delta}\|\boldsymbol{x}_{\text{orig}}\|_2^2
\end{aligned}
\tag{59}
$$

□

B Additional Plots

B.1 Wine data set

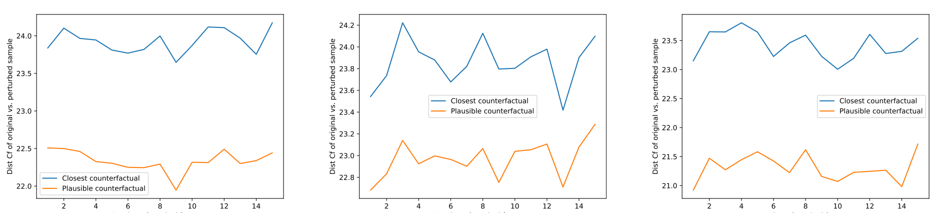
Fig. 3: Wine data set: Median absolute distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (8)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression Decision tree GLVQ

B.2 Breast cancer data set

Fig. 4: Breast cancer data set: Median absolute distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (8)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression Decision tree GLVQ