Evaluating Robustness of Counterfactual Explanations

1st André Artelt
CITEC
Bielefeld University
Bielefeld, Germany
aartelt@techfak.de

2nd Valerie Vaquet
CITEC
Bielefeld University
Bielefeld, Germany
vvaquet@techfak.de

3rd Riza Velioglu
CITEC
Bielefeld University
Bielefeld, Germany
rvelioglu@techfak.de

4th Fabian Hinder
CITEC
Bielefeld University
Bielefeld, Germany
fhinder@techfak.de

5th Johannes Brinkrolf
CITEC
Bielefeld University
Bielefeld, Germany
jbrinkro@techfak.de

6th Malte Schilling
CITEC
Bielefeld University
Bielefeld, Germany
mschilli@techfak.de

7th Barbara Hammer
CITEC
Bielefeld University
Bielefeld, Germany
bhammer@techfak.de

***Abstract*—Transparency is a fundamental requirement for decision making systems when these should be deployed in the real world. It is usually achieved by providing explanations of the system's behavior. A prominent and intuitive type of explanations are counterfactual explanations. Counterfactual explanations explain a behavior to the user by proposing actions—as changes to the input—that would cause a different (specified) behavior of the system. However, such explanation methods can be unstable with respect to small changes to the input—i.e. even a small change in the input can lead to huge or arbitrary changes in the output and of the explanation. This could be problematic for counterfactual explanations, as two similar individuals might get very different explanations. Even worse, if the recommended actions differ considerably in their complexity, one would consider such unstable (counterfactual) explanations as individually unfair.**

In this work, we formally and empirically study the robustness of counterfactual explanations in general, as well as under different models and different kinds of perturbations. Furthermore, we propose that plausible counterfactual explanations can be used instead of closest counterfactual explanations to improve the robustness and consequently the individual fairness of counterfactual explanations.

***Index Terms*—XAI, Contrasting Explanations, Counterfactual Explanations, Robustness, Fairness**

I. Introduction

Transparency is a fundamental building block for ethical artificial intelligence (AI) and machine learning (ML) based decision making systems. In particular, the increasing use of automated decision making systems in the real world [1], [2] has strengthened the demand for trustworthy systems. The criticality of transparency was also recognized recently by policy makers which resulted in legal regulations like the EU's GDPR [3] 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 [4]–[7]. Nowadays, there are several technologies available for explaining ML models [5], [8]. One specific family of methods are model-agnostic methods [5], [9], which are not tailored to a particular model or representation and thus applicable to a wide range of ML models.

Examples of model-agnostic methods are feature interaction methods [10], feature importance methods [11], partial dependency plots [12], and methods that approximate the model locally by an explainable model [13], [14]. These methods explain the models by using *input features* as the vocabulary.

In contrast, in *example-based explanations* [15] a prediction or behavior is explained by a (set of) data points. Instances of example-based explanations are prototypes & criticisms [16], influential instances [17], and counterfactual explanations [18]. A counterfactual explanation states a change to the original input that results in a different (specific) prediction or behavior of the decision making system— *what needs to be different in order for the system's prediction to change?* 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* [8], [18]. Furthermore, there exist strong evidence that explanations by humans are often counterfactual in nature [19] and preferred. Our focus will be on counterfactual explanations.

However, it has also been shown that many explanation methods are vulnerable to adversarial attacks [20]–[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 that should be explained. This instability of explanations applies to counterfactual explanations as well [23]. As the necessity of local stability of explanations is widely

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*, fundings from the federal state government of North Rhine-Westphalia (NRW) for the projects *Bias von KI-Modelle bei der Informationsbildung und deren Implikationen in der Wirtschaft* and the research training group *Dataninja* (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis), and funding from the German Federal Ministry of Education and Research (BMBF) through the project *TiM* (05M20PBA).

accepted [20], [21], [23]–[25], we face the challenge how to compute stable and robust counterfactual explanations which is still an open-research problem [26].

Missing stability and robustness of explanations can lead to unfair explanations and thus compromise the system's trustworthiness [20], [25]. In particular, this holds true for counterfactual explanations because their explanations lead to actions: since these can be interpreted as proposed actions in order to achieve a desired goal, it is problematic, if the proposed actions differ considerably between individuals without a clear reason. *For instance consider the scenario of loan application where two (financially) similar individuals applied for a loan and both applications got rejected. However, the counterfactual explanation both people get are highly different—one applicant only needs a minor increase in monthly income, while the other applicant is told that a major increase in monthly income as well as some other aspects are required for getting the loan application accepted.* Thus, robustness of counterfactual explanations is closely related to individual fairness of counterfactual explanations.

Note that fairness of ML based decision making systems itself has been already studied for some time. Especially because of prominent failures of ML system in critical situations like predictive policing [1] and loan approval [2]. Many of those failures, where the systems exhibit unethical behavior, deal with predictions that are correlated to sensitive attributes such as race, gender, etc. As a consequence, several (formal) fairness criteria, concerning the predictions and behaviors of ML systems, have been proposed [27], [28]. A large number of these criteria deal with the dependency between a sensitive attribute/feature and the response variable (prediction/behavior of the system)—these are known as group-based fairness criteria. Prominent examples are [28] demographic parity, equalized odds, predictive rate parity, or causal independence. However, there are 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 focuses on individual (rather than group) fairness [29]. The idea behind individual fairness is to treat similar individuals similarly—this notion resembles the concept of individual fairness from other scientific disciplines [29], [30]. Despite its appealing intuition, a major challenge of individual fairness constitutes in its formalization—i.e. given two individuals, how can we score their mutual similarity?

Despite its relevance, robustness of explanations has only recently started being investigated in first approaches [22], [31], although it is well known that robustness and fairness concerning the predictions and behavior of ML models are closely related to each other [32], [33]. Consequently, only a few papers study the fairness of explanations itself. For instance the authors of [34] 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 [35] 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 propose a formalization of robustness of counterfactual explanations and 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 plausible counterfactual are more robust. Furthermore, we also relate our robustness definition to individual fairness and argue that our robustness definition can be interpreted as a definition of individual fairness and thus all our findings on robustness transfer to individual fairness of counterfactual explanations as well.

To address these goals, first, we briefly review counterfactual explanations in Section II. Next, we formally define our notion of robustness of counterfactual explanations in Section III-A. In Section III-C, we study formal properties of robustness of counterfactual explanations (Section III-C), we propose the concept to use plausible instead of closest counterfactual explanations for improving the robustness of the explanations and also show how our robustness definition relates to individual fairness of counterfactual explanations (Section III-E). We empirically evaluate the robustness of closest and plausible counterfactual explanations in Section IV. Finally, we give a summary and outlook in Section V.

II. Foundations

A. Counterfactual Explanations

Counterfactual explanations (often just called counterfactuals) contrast samples by counterparts with minimum change of the appearance but different class label [8], [18] and can be formalized as follows:

Definition 1 ((Closest) Counterfactual explanation [18])**.** *Assume a prediction function $h : \mathbb{R}^d \to \mathcal{Y}$ is given. Computing a counterfactual $\vec{x}_{cf} \in \mathbb{R}^d$ for a given input $\vec{x}_{orig} \in \mathbb{R}^d$ is phrased as an optimization problem:*

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

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

In this work we assume that data comes from a real-vector space and continuous optimization is possible. In this context [36], 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 $\vec{x}_{\text{orig}}$ and desired label y', the counterfactual provided by the algorithm at hand is referred to as $\vec{x}_{\text{cf}} = \text{CF}(\vec{x}_{\text{orig}}, y')$.

While the formalization of the optimization problem Eq. (1) is model agnostic (i.e. it does not make any assumptions on the model $h(\cdot)$), it can be beneficial to rewrite the optimization problem Eq. (1) in constrained form [36]:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \theta(\vec{x}_{\text{cf}}, \vec{x}_{\text{orig}}) \tag{2a}$$

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

The authors of [36] 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 [37], the constrained form [36] becomes superior over the original black-box modelling [18] if we have access to the underlying model $h(\cdot)$. Furthermore, we can also integrate affine pre-processings like PCA into the convex programs and therefore still compute counterfactuals in the original data space [36].

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 [38]–[40]. To overcome this problem, several approaches [38], [39] 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 [38] build upon the constrained form Eq. (2) and propose the following extension of Eq. (2) for computing plausible counterfactuals:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \theta(\vec{x}_{\text{cf}}, \vec{x}_{\text{orig}}) \tag{3a}$$

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

$$\hat{p}_{y'}(\vec{x}_{\text{cf}}) \geq \delta \tag{3c}$$

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 be phrased as a set of convex quadratic constraints [36].

Unstable counterfactuals can be considered as 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 propose a formalization of robustness of counterfactual explanations.

III. Robustness of Counterfactual Explanations

We are not interested in robustness 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 robustness of counterfactual explanations in general, and investigate its specific properties for important specific prediction functions $h : \mathcal{X} \to \mathcal{Y}$.

A. Formalization of Robustness

We aim for a formalization of robustness of counterfactual explanations. Inspired by the intuition (and formalization) of robustness of explanation methods [22]), we propose the following definition for locally measuring the influence of small perturbations of the input to the counterfactual explanation:

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

Let $\vec{x}$ be a perturbed sample of $\vec{x}_{orig}$ that is $\vec{x} \sim p_\epsilon(\vec{x}_{orig})$ where $p_\epsilon(\cdot)$ denotes the density of the perturbed samples.Denote $y = y'$ if $h(\vec{x}) = y_{orig}$ and $y = y_{orig}$ otherwise. Let $\vec{x}'_{cf} = CF(\vec{x}, y) \in \mathcal{X}$ be a counterfactual explanations of this perturbed sample $(\vec{x}, y_{orig})$.

Given a function $\mathrm{d}(\cdot)$ for computing the similarity/distance between two given counterfactual explanations, we measure the amount of stability of the explanation $\vec{x}_{cf}$ as the expected similarity/distance between the counterfactual explanations of the original sample $\vec{x}_{orig}$ and such a perturbed sample $\vec{x}$:

$$\mathop{\mathbb{E}}_{\vec{x} \sim p_\epsilon(\vec{x}_{orig})} \left[\mathrm{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \right] \tag{4}$$

assuming that Eq. (4) *exists and is well defined.*

Note that the instability Eq. (4) is to be minimized - i.e. smaller values correspond to a better robustness.

B. Natural Perturbations

While there are different possible ways of perturbing a given input $\vec{x}_{\text{orig}}$ - i.e. choosing $p_\epsilon(\cdot)$ 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.

1) Gaussian Noise: Perturbing a given input $\vec{x}_{\text{orig}}$ with Gaussian noise means to add a small amount of normally distributed noise $\vec{\delta}$ to $\vec{x}_{\text{orig}}$:

$$\vec{x} = \vec{x}_{\text{orig}} + \vec{\delta} \quad \text{where } \vec{\delta} \sim \mathcal{N}(\vec{0}, \mathbf{\Sigma}) \tag{5}$$

we therefore define:

$$p_\epsilon(\vec{x}_{\text{orig}}) = \mathcal{N}(\vec{x}_{\text{orig}}, \mathbf{\Sigma}) \tag{6}$$

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 cannot guarantee that the perturbed sample stays close to the original sample $\vec{x}_{\text{orig}}$ because $\mathcal{N}(\vec{0}, \mathbf{\Sigma})$ can

yield arbitrarily large values. The probability of such values is limited, however.

2) Bounded Uniform Noise: Similar to a perturbation with Gaussian noise (see Eq. (5)), we add a small amount of uniformly distributed noise $\vec{\delta}$ to $\vec{x}_{\text{orig}}$:

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

we therefore define:

$$p_\epsilon(\vec{x}_{\text{orig}}) = \mathcal{U}(-\epsilon\vec{x}_{\text{orig}}, \epsilon\vec{x}_{\text{orig}}) \tag{8}$$

where $\epsilon > 0$ controls and upper bounds the amount of noise - note that in Eq. (7) we use the same ϵ for every dimension, however using different ϵ for each dimension is also possible.

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

3) Feature Masking: While a perturbation by Gaussian noise Eq. (5) or bounded uniform noise Eq. (7)[1] potentially changes every feature, feature masking allows a more precise way of perturbing a given input:

$$\vec{x} = \vec{x}_{\text{orig}} \odot \vec{m} \quad \text{where } \vec{m} \in \{0, 1\}^d \tag{9}$$

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. Note that in comparison to Gaussian and uniform noise, this is not a proper density and how close the perturbed sample stays to the original sample completely depends on the chosen mask.

C. Analyzing Robustness

In this section we formally study robustness of closest counterfactuals under different prediction functions $h(\cdot)$. First, we give a (rather trivial) 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 $(\vec{x}_{orig}, y_{orig}) \in \mathcal{X} \times \mathcal{Y}$ be a sample for which we are given a closest counterfactual (see Eq. (2)) $\vec{x}_{cf} = CF(\vec{x}_{orig}, y') \in \mathcal{X}$. Let $\vec{x} \in \mathcal{X}$ be a perturbed version of $\vec{x}_{orig}$ such that $\|\vec{x}_{orig} - \vec{x}\|_p \leq \epsilon$ – we denote the corresponding closest counterfactual of $\vec{x}$ as $\vec{x}'_{cf}$. We can then bound the difference between the two counterfactuals $\vec{x}_{cf}$ and $\vec{x}'_{cf}$ as follows:*

$$\|\vec{x}_{cf} - \vec{x}'_{cf}\|_p \leq 2\epsilon + 2\|\vec{x}_{orig} - \vec{x}_{cf}\|_p \tag{10}$$

In case of a binary linear classifier, we can refine the bound from Theorem 1 as stated in Corollary 2.

[1]Although one could use different (possible empty) intervals for different dimensions.

Corollary 2 (Bound on closest counterfactuals of perturbed samples for binary linear classifiers)**.** *In case of a binary linear classifier, i.e. $h(\vec{x}) = \text{sign}\left(\vec{w}^\top \vec{x}\right)$ with $\|\vec{w}\|_p = 1$, we can refine the bound from Theorem 1 as follows:*

$$\|\vec{x}_{cf} - \vec{x}'_{cf}\|_p \leq 2\epsilon + 2|\vec{w}^\top \vec{x}_{orig}| \tag{11}$$

Remark 1. *Note that while the general bound Eq. (10) in Theorem 1 depends on the closest counterfactual of the unperturbed sample $\vec{x}_{orig}$, the bound Eq. (11) in Corollary 2 only depends on the original sample $\vec{x}_{orig}$, the model parameter $\vec{w}$ and the perturbation bound ϵ. Both bounds (Theorem 1 and Corollary 2) are rather loose because they do not make any assumption on the perturbation $p_\epsilon(\cdot)$ except that it must be bounded by ϵ - in addition the bound in Theorem 1 does not even make any assumption on $h(\cdot)$ at all.*

1) Gaussian Noise: Additional assumptions allow more precise (and potentially more useful) statements as shown in the next Theorem 3. There (and the consequential Corollary 4 and Corollary 5) we study the robustness of closest counterfactual explanations of a binary linear classifier under Gaussian noise. It turns out, that the robustness 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 instability.

Theorem 3 (Instability 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. $h(\vec{x}) = \text{sign}\left(\vec{w}^\top \vec{x}\right)$ with $\|\vec{w}\|_2 = 1$. The instability of closest counterfactuals (see Definition 2) under Gaussian noise Eq. (5) - with an arbitrary diagonal covariance $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$ - at a sample $(\vec{x}_{orig}, y_{orig}) \in \mathbb{R}^d \times \{-1, 1\}$ can be stated as follows:*

$$\mathop{\mathbb{E}}_{\vec{x} \sim \mathcal{N}(\vec{x}_{orig}, \mathbf{\Sigma})} \left[\text{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \right] = \text{trace}(\mathbf{\Sigma}) - \vec{w}^\top \mathbf{\Sigma} \vec{w} \tag{12}$$

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

Corollary 4. *If we assume the identity matrix $\mathbb{I}$ as a covariance matrix $\mathbf{\Sigma}$ of the Gaussian noise in Theorem 3, Eq. (12) simplifies as follows:*

$$\mathop{\mathbb{E}}_{\vec{x} \sim \mathcal{N}(\vec{x}_{orig}, \mathbb{I})} \left[\text{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \right] = d - 1 \tag{13}$$

Corollary 5. *In the setting of Theorem 3 and Corollary 4, the probability that the difference is larger than a given $\delta > 0$ is bounded as follows:*

$$\mathbb{P}\Big(\text{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \geq \delta \Big) \leq \frac{d - 1}{\delta} \tag{14}$$

Remark 2. *We can interpret Theorem 3 (and in particular the consequential Corollary 4 and Corollary 5) as a "curse of dimensionality" for robustness of closest counterfactual explanations: the larger the dimension d of the data space, the larger the potential instability if the data manifold is intrinsically locally high dimensional.*

2) Bounded Uniform Noise: We can still make some statements on the robustness Definition 2 of closest counterfactual explanations, when using bounded uniform noise instead of Gaussian noise, as stated in Theorem 6 and Corollary 7.

Theorem 6 (Instability 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. $h(\vec{x}) = \text{sign}\left(\vec{w}^\top \vec{x}\right)$ with $\|\vec{w}\|_2 = 1$. The instability of closest counterfactuals (see Definition 2) under a bounded uniform noise Eq.* (7) *at an arbitrary sample $(\vec{x}_{orig}, y_{orig}) \in \mathbb{R}^d \times \{-1, 1\}$ can be stated as follows:*

$$\mathop{\mathbb{E}}_{\vec{x} \sim \mathcal{U}(\vec{x}_{orig} \pm \epsilon \vec{1})} \left[\mathrm{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \right] = \frac{\epsilon^2 (d-1)}{3} \tag{15}$$

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

Corollary 7. *In the setting of Theorem 6, the probability that the instability is larger than some $\delta > 0$ can be upper bounded as:*

$$\mathbb{P}\Big(\mathrm{d}(\vec{x}_{cf}, \vec{x}'_{cf}) \geq \delta \Big) \leq \frac{\delta \epsilon^2 (d-1)}{3} \tag{16}$$

D. Discussion

We will empirically confirm these mathematical findings of a potential presence of instability even for simple models in the experiments (see Section IV). As a mediation, we propose to add a regularization for improving the robustness Definition 2 of counterfactual explanations. One solution can exist in the reference to plausible counterfactuals instead of closest ones. One reason of instability 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.

E. Relation to Individual Fairness

Recall that, while many fairness criterions are concerned with protected groups, individual fairness focuses on individuals and requires to "treat similar individuals similarly" [29]. Given a prediction function $h : \mathcal{X} \to \mathcal{Y}$, we can formalize individual fairness as follows:

$$\mathrm{d}(\vec{x}_1, \vec{x}_2) \leq \epsilon_1 \implies \Delta\big(h(\vec{x}_1), h(\vec{x}_2)\big) \leq \epsilon_2 \qquad \forall\, \vec{x}_1, \vec{x}_2 \in \mathcal{X} \tag{17}$$

where $\mathrm{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denotes a similarity measure on the individuals in $\mathcal{X}$, $\epsilon_1, \epsilon_2 > 0$ denotes a threshold up to which we consider two individuals / the predictions as similar and $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denotes a similarity measure on the predictions of the ML system $h(\cdot)$. A critical choice, which highly depends on the specific use-case, are the similarity measures $\mathrm{d}(\cdot)$ and $\Delta(\cdot)$. A common default choice is a p-norm assuming real-vector spaces like we do in this work.

As already mentioned in the introduction of this work, robustness and (individual) fairness are highly related to each other. In particular, we can interpret our definition of local instability of counterfactual explanation (Definition 2) as a measure for individual fairness of counterfactual explanations - i.e. the larger the instability, the larger the individual unfairness of the counterfactual explanation. Therefore, our robustness studies can also be interpreted as a study of individual fairness. Note that in contrast to the definition of individually fair causal recourse as proposed in [35], 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).

IV. Experiments

We empirically evaluate the robustness (Definition 2) of closest and plausible counterfactual explanations. Furthermore, we also empirically confirm the occurrence of the "curse of dimensionality" (see Remark 2) 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.

a) 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 2).
- The "Breast Cancer Wisconsin (Diagnostic) Data Set" [41] whereby we add a PCA dimensionality reduction to 5 dimensions to the model.
- The "Wine data set" [42].
- The "Optical Recognition of Handwritten Digits Data Set" [43] whereby we add a PCA dimensionality reduction to 40 dimensions to the model.

b) 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 for all decision trees we set the maximum depth of each tree to 7.

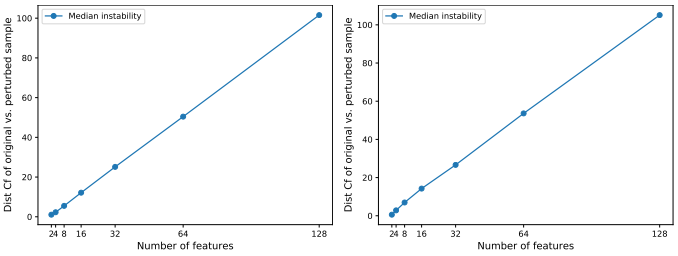
c) Curse of Dimensionality: In order to empirically study the occurrence of the "curse of dimensionality" (see Remark 2), we use the Gaussian blobs toy data set for fitting and evaluating counterfactuals (original vs. perturbed sample) under a decision tree classifier and a GLVQ model. The results (over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

d) 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

TABLE I: Comparing the median l_1 distance between counterfactual of original sample and perturbed sample (using Gaussian noise Eq. (5)) - closest and plausible counterfactual explanations. Smaller values are better - best values are **highlighted**.

Data set	Wine		Breast cancer		Handwritten digits	
Method	Closest	Plausible	Closest	Plausible	Closest	Plausible
Softmax	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**

Fig. 1: Gaussian blobs: Median l_1 distance between counterfactual of original sample and perturbed sample (using Gaussian noise Eq. (5) with $\mathbf{\Sigma} = \mathbb{I}$). Smaller values are better.



Decision tree GLVQ

original as well as the perturbed sample. We use two different types of perturbations: Gaussian noise Eq. (5) with $\mathbf{\Sigma} = \mathbb{I}$ and feature masking Eq. (9) 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. (4) - we do this separately for closest and plausible counterfactuals. Furthermore, we use MOSEK[2] as a solver for all mathematical programs. The complete implementation of the experiments is available on GitHub[3].

e) Results: The results of using Gaussian noise Eq. (5) for perturbing the samples are shown in Table I. The results on the digit data set for increasingly masking more and more features Eq. (9) are shown in Fig. 2 - plots for the other data sets are given in appendix A.

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 more robust (Definition 2). The size of the differences depends a lot on the combination of model and data set. However, in all cases there is a clear difference. 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 plausible counterfactuals are more robust than the closest counterfactuals - even when masking up to 50% of all features.

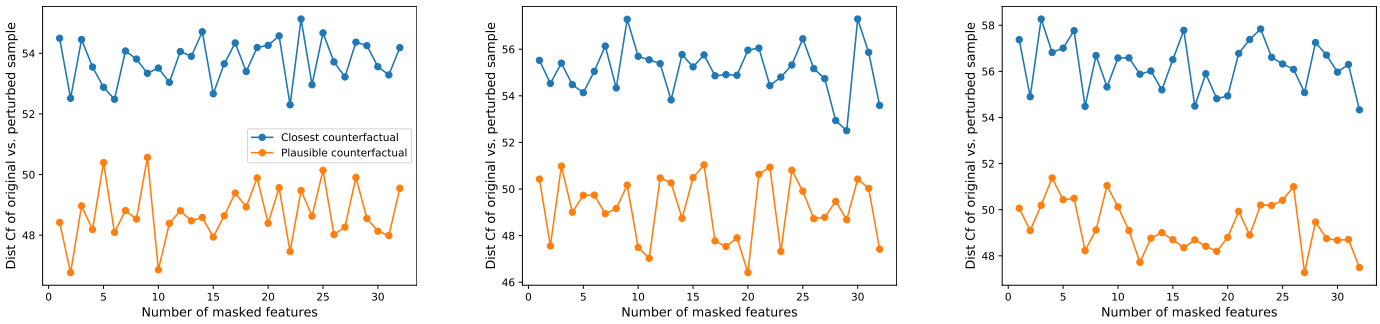
V. Discussion and Conclusion

In this work, we studied the robustness of counterfactual explanations. We found that closest counterfactuals can be unstable in the sense that they are sensitive to different kinds of small perturbations. We proposed to use plausible instead of closest counterfactuals for increasing the robustness of counterfactual explanations which we empirically evaluated providing a comparison of the robustness of closest vs. plausible counterfactual explanations. We found evidence that plausible counterfactuals provide better robustness than closest counterfactual explanations. We also argued that robustness and individual fairness of counterfactual explanations are basically the same and thus, our findings on robustness also apply to individual fairness of counterfactual explanations — i.e. the individual fairness of closest counterfactual explanation is rather poor and using plausible counterfactuals yield a better individual fairness.

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 [38] we used in this article is only one possible approach for computing plausible counterfactuals, for other approaches see [39], [40]. 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/modeled by our proposed formalization and methods.

[2]We gratefully acknowledge an academic license provided by MOSEK ApS.
[3]https://github.com/andreArtelt/FairnessRobustnessContrastingExplanations

Fig. 2: Handwritten digits data set: Median l_1 distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (9)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression | Decision tree | GLVQ

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias - theres software used across the country to predict future criminals. and its biased against blacks." 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] K. Waddell, "How algorithms can bring down minorities' credit scores," The Atlantic, 2016.

[3] E. parliament and 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.

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "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, 2018, pp. 80–89. [Online]. Available: https://doi.org/10.1109/DSAA.2018.00018

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018. [Online]. Available: http://doi.acm.org/10.1145/3236009

[6] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): towards medical XAI," CoRR, vol. abs/1907.07374, 2019. [Online]. Available: http://arxiv.org/abs/1907.07374

[7] W. Samek, T. Wiegand, and K. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," CoRR, vol. abs/1708.08296, 2017. [Online]. Available: http://arxiv.org/abs/1708.08296

[8] C. Molnar, Interpretable Machine Learning, 2019, https://christophm.github.io/interpretable-ml-book/.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," in ICML Workshop on Human Interpretability in Machine Learning (WHI), 2016.

[10] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," CoRR, vol. abs/1805.04755, 2018. [Online]. Available: http://arxiv.org/abs/1805.04755

[11] A. Fisher, C. Rudin, and F. Dominici, "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, p. arXiv:1801.01489, Jan 2018.

[12] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," Journal of Business & Economic Statistics, vol. 0, no. ja, pp. 1–19, 2019. [Online]. Available: https://doi.org/10.1080/07350015.2019.1624293

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 1135–1144. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939778

[14] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," CoRR, vol. abs/1805.10820, 2018. [Online]. Available: http://arxiv.org/abs/1805.10820

[15] A. Aamodt and E. Plaza., "Case-based reasoning: Foundational issues, methodological variations, and systemapproaches." AI communications, 1994.

[16] B. Kim, O. Koyejo, and R. Khanna, "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, 2016, pp. 2280–2288.

[17] P. W. Koh and P. Liang, "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, 2017, pp. 1885–1894. [Online]. Available: http://proceedings.mlr.press/v70/koh17a.html

[18] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," CoRR, vol. abs/1711.00399, 2017. [Online]. Available: http://arxiv.org/abs/1711.00399

[19] R. M. J. Byrne, "Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 6276–6282. [Online]. Available: https://doi.org/10.24963/ijcai.2019/876

[20] C. J. Anders, P. Pasliev, A. Dombrowski, K. Müller, and P. Kessel, "Fairwashing explanations with off-manifold detergent," in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 314–323. [Online]. Available: http://proceedings.mlr.press/v119/anders20a.html

[21] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 2921–2932. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html

[22] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," CoRR, vol. abs/1806.08049, 2018. [Online]. Available: http://arxiv.org/abs/1806.08049

[23] T. Laugel, M. Lesot, C. Marsala, and M. Detyniecki, "Issues with post-hoc counterfactual explanations: a discussion," CoRR, vol. abs/1906.04774, 2019. [Online]. Available: http://arxiv.org/abs/1906.04774

[24] T. Laugel, X. Renard, M. Lesot, C. Marsala, and M. Detyniecki, "Defining locality for surrogates in post-hoc interpretablity," CoRR, vol. abs/1806.07498, 2018. [Online]. Available: http://arxiv.org/abs/1806.07498

[25] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7786–7795. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[26] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," 2020.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," CoRR, vol. abs/1908.09635, 2019. [Online]. Available: http://arxiv.org/abs/1908.09635

[28] S. Caton and C. Haas, "Fairness in machine learning: A survey," CoRR, vol. abs/2010.04053, 2020. [Online]. Available: https://arxiv.org/abs/2010.04053

[29] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," in Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012, S. Goldwasser, Ed. ACM, 2012, pp. 214–226. [Online]. Available: https://doi.org/10.1145/2090236.2090255

[30] R. Binns, "On the apparent conflict between individual and group fairness," in FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, Eds. ACM, 2020, pp. 514–524. [Online]. Available: https://doi.org/10.1145/3351095.3372864

[31] L. Hancox-Li, "Robustness in machine learning explanations: Does it matter?" ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 640647. [Online]. Available: https://doi.org/10.1145/3351095.3372836

[32] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," CoRR, vol. abs/2006.08669, 2020. [Online]. Available: https://arxiv.org/abs/2006.08669

[33] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dicker-son., "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. [Online]. Available: https://doi.org/10.1145/3442188.3445910

[34] V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian, "Equalizing recourse across groups," CoRR, vol. abs/1909.03166, 2019. [Online]. Available: http://arxiv.org/abs/1909.03166

[35] J. von Kgelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schlkopf, "On the fairness of causal algorithmic recourse," 2021.

[36] A. Artelt and B. Hammer, "On the computation of counterfactual explanations - A survey," CoRR, vol. abs/1911.07749, 2019. [Online]. Available: http://arxiv.org/abs/1911.07749

[37] S. Boyd and L. Vandenberghe, Convex Optimization. New York, NY, USA: Cambridge University Press, 2004.

[38] A. Artelt and B. Hammer, "Convex density constraints for computing plausible counterfactual explanations," 29th International Conference on Artificial Neural Networks (ICANN), 2020.

[39] A. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," CoRR, vol. abs/1907.02584, 2019.

[40] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. D. Bie, and P. A. Flach, "FACE: feasible and actionable counterfactual explanations," CoRR, vol. abs/1909.09369, 2019. [Online]. Available: http://arxiv.org/abs/1909.09369

[41] O. L. M. William H. Wolberg, W. Nick Street, "Breast cancer wisconsin (diagnostic) data set," https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), 1995.

[42] D. C. S. Aeberhard and O. de Vel, "Comparison of classifiers in high dimensional settings," Tech. Rep. no. 92-02, 1992.

[43] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits data set," https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits, 1998.

APPENDIX

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

$$\|\vec{x}_{\text{orig}} - \vec{x}\|_{\text{p}} \leq \epsilon \tag{18}$$

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

$$\|\vec{x} - \vec{x}'_{\text{cf}}\|_{\text{p}} \leq \|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} \tag{19}$$

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

$$\|\vec{x}'_{\text{cf}} - \vec{x}_{\text{cf}}\|_{\text{p}} \leq \|\vec{x} - \vec{x}'_{\text{cf}}\|_{\text{p}} + \|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} \tag{20}$$

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

$$\begin{aligned} \|\vec{x}_{\text{cf}} - \vec{x}'_{\text{cf}}\|_{\text{p}} &\leq \|\vec{x} - \vec{x}'_{\text{cf}}\|_{\text{p}} + \|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} \\ &\leq \|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} + \|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} \\ &= 2\|\vec{x} - \vec{x}_{\text{cf}}\|_{\text{p}} \end{aligned} \tag{21}$$

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

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

Plugging Eq. (22) into Eq. (21) yields the desired bound Eq. (10):

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

■

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

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

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

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

$$\text{s.t. } \vec{w}^\top \vec{x}_{\text{cf}} = 0 \tag{25b}$$

Note that the constraint Eq. (25b) "replaces/approximates" the constraint $h(\vec{x}_{\text{cf}}) = y'$ Eq. (2b). The constraint Eq. (25b) requires that the solution $\vec{x}_{\text{cf}}$ 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:

$$\mathcal{L}(\vec{x}_{\text{cf}}, \lambda) = \vec{x}_{\text{orig}}^\top \vec{x}_{\text{orig}} - 2\vec{x}_{\text{orig}}^\top \vec{x}_{\text{cf}} + \vec{x}_{\text{cf}}^\top \vec{x}_{\text{cf}} - \lambda \vec{w}^\top \vec{x}_{\text{cf}} \quad (26)$$

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

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

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

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

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

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

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

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

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

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

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

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

which concludes this sub-proof.
Plugging Eq. (24) into the bound Eq. (10) from Theorem 1, and again assuming w.l.o.g. that $\|\vec{w}\|_2 = 1$, yields the desired bound Eq. (11):

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

■

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

$$\vec{x}_{\text{cf}} = \vec{x} - \left(\vec{w}^\top \vec{x}\right)\vec{w} \quad (34)$$

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

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

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

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

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

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

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

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

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

where we made use of the assumption that $\|\vec{w}\|_2 = 1$. Substituting Eq. (37), Eq. (38), Eq. (39), Eq. (40) in Eq. (36) yields:

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

which concludes the proof. ■

4) *Corollary 4:* Substituting $\mathbb{I}$ for $\mathbf{\Sigma}$ in Eq. (12) from Theorem 3 yields the claimed expectation:

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

■

5) *Corollary 5:* Plugging the expectation from Corollary 4 into Markov's inequality yields the claimed bound:

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

■

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

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

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

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

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

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

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

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

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

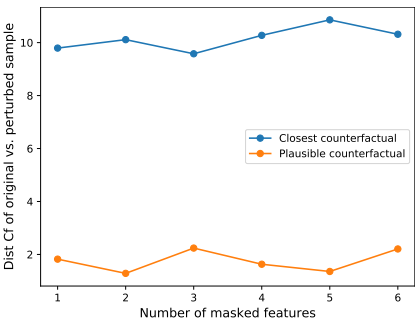
which concludes the proof. ■

7) *Corollary 7:* Plugging the expectation from Theorem 6 into Markov's inequality yields the claimed bound:

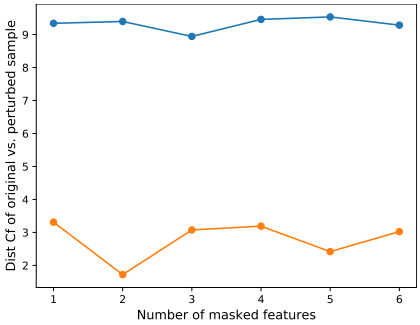
$$\begin{aligned}
\mathbb{P}\Big(\,\mathrm{d}(\vec{x}_{\text{cf}},\vec{x}'_{\text{cf}}) \geq \delta\Big) &\leq \frac{\mathbb{E}\big[\,\mathrm{d}(\vec{x}_{\text{cf}},\vec{x}'_{\text{cf}})\big]}{\delta} \\
&= \frac{\delta\epsilon^2(d-1)}{3}
\end{aligned} \tag{50}$$

■

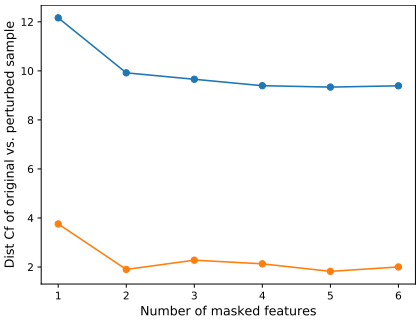
Fig. 3: Wine data set: Median l_1 distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (9)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression

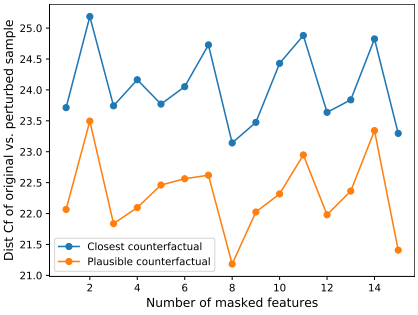


Decision tree

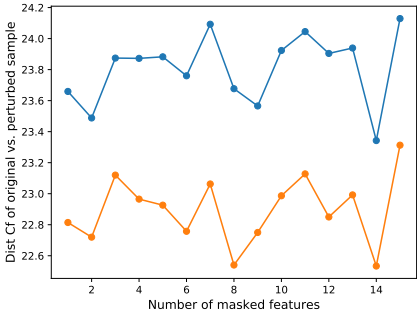


GLVQ

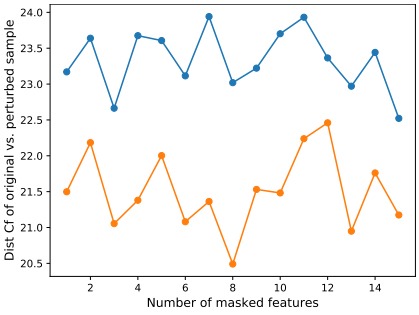
Fig. 4: Breast cancer data set: Median l_1 distance between counterfactual of original sample and perturbed sample (using feature masking Eq. (9)) for closest and plausible counterfactual explanations - for different number of masked features. Smaller values are better.



Softmax regression



Decision tree



GLVQ