On Counterfactual Explanations under Predictive Multiplicity

Martin Pawelczyk
University of Tuebingen
Tuebingen, Germany

Klaus Broelemann
Schufa AG
Wiesbaden, Germany

Gjergji Kasneci
University of Tuebingen
Tuebingen, Germany

Abstract

Counterfactual explanations are usually obtained by identifying the *smallest change* made to an input to change a prediction made by a fixed model (hereafter called *sparse methods*). Recent work, however, has revitalized an old insight: there often does not exist one superior solution to a prediction problem with respect to commonly used measures of interest (e.g. error rate). In fact, often multiple different classifiers give almost equal solutions. This phenomenon is known as *predictive multiplicity* (Breiman, 2001; Marx et al., 2019). In this work, we derive a general upper bound for the costs of counterfactual explanations under predictive multiplicity. Most notably, it depends on a *discrepancy* notion between two classifiers, which describes how differently they treat negatively predicted individuals. We then compare *sparse* and *data support* approaches empirically on real-world data. The results show that *data support* methods are more robust to multiplicity of different models. At the same time, we show that those methods have provably higher cost of generating counterfactual explanations under one fixed model. In summary, our theoretical and empirical results challenge the commonly held view that counterfactual recommendations should be *sparse* in general.

1 INTRODUCTION

Counterfactual explanations are usually obtained by identifying the *smallest change* made to an input vector to qualitatively influence a prediction of a pretrained classifier in a positive way; for example, from 'loan rejected' to 'awarded' or from 'high risk of cardiovascular disease' to 'low risk'. But what is a good counterfactual?

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

A tale of 2 camps. The literature commonly agrees that counterfactual explanations mainly serve two purposes (Ustun et al., 2019; Karimi et al., 2020a; Wachter et al., 2017): First, they should help understand a model's *local decision boundary*, answering questions like *"Why does the model give a certain prediction for a given individual?"* (*purpose I*). Second, counterfactual explanations should provide *a recommendation/recommendations* for the individual in question (*purpose II*). Hence, they should give answers to the question *"What is the* smallest change *in inputs an individual needs to make in the future to receive a desired outcome?"*. In this work, we focus on purpose II and analyze counterfactual explanations from the recommendation perspective.

Often there exist several ways to make a reasonable recommendation. So, what constitutes a reasonable recommendation? We roughly split the current literature into two camps. The first line of work (we call them *sparse* counterfactuals) assumes that counterfactual recommendations with minimal change in ℓ_p-norm are most desirable (Wachter et al., 2017; Grath et al., 2018; Russell, 2019; Ustun et al., 2019; Laugel et al., 2017; Karimi et al., 2020a; Tolomei et al., 2017).

The second camp (henceforth called *(Data) Support* counterfactuals) suggests that the norm should receive second order importance when generating counterfactual explanations. Instead, it would be more desirable to generate counterfactual recommendations that are close to correctly classified observations from the desired class and semantically meaningful (Laugel et al., 2019b,a; Pawelczyk et al., 2020; Joshi et al., 2019; Mahajan et al., 2019). We give exact definitions of both types in section 2. Finally, a recent line of work considers causal interventions to generate counterfactual explanations (Karimi et al., 2020b).

All aforementioned works assume that the pretrained classifier is given and that there exist no uncertainty as to whether it is the best possible classifier or whether it will remain the classifier of choice over time. Counterfactual recommendations are then usually generated with reference to this 'best' pretrained model.

Proposal	Input subset	current value		required
1	`#credit cards`	5	→	3
2	`current debt`	\$3250	→	\$1000
3	`has savings account`	0	→	1
	`has retirement account`	0	→	1

Table 1: Stylised example from an individual who was denied credit by a fixed classifier f, i.e. $\text{sign}(f(\boldsymbol{x}_{current})) = -1$, and three different associated counterfactual recommendations, i.e. $\text{sign}(f(\boldsymbol{x}_{required})) = +1$. **The difference between the current values and the required values are the costs of counterfactual recommendations**. Example taken from Ustun et al. (2019).

Counterfactuals under model multiplicity. Recent work has revitalized an old insight (Marx et al., 2019): there often does not exist one superior solution to a prediction problem with respect to commonly used measures of interest (e.g. error rate). In fact, often multiple different models give almost equal solutions. This phenomenon is known as *predictive multiplicity* (Marx et al., 2019; Breiman, 2001; McCullagh and Nelder, 1989) and in this work we argue that it should shape our understanding of how counterfactual recommendations are generated.

Admitting the existence of several well performing models for the very same prediction task calls the entire business of generating counterfactual recommendations for one particular model into question. Or, as Leo Breiman already put it (Breiman, 2001): *"[...] if there exist several equally good models for a given dataset [sic], each of which provides a different explanation of the data-generating process, then how can we tell which one is correct?"*.

Barocas et al. (2020) identify 4 hidden assumptions underlying the generation of counterfactual explanations: (**A1**) *The underlying model is stable over time*. (**A2**) *Explanations can be offered without regard to decision making in other areas of people's lives*. (**A3**) *Counterfactual explanations map to real world actions*. (**A4**) *Inputs can be made commensurate by looking at the training data*. In this work, we will investigate the effect of assumption **A1** for counterfactual explanations in theory and in practice.

To get a better understanding for the underlying problem, let us consider the two following scenarios:

(a) The decision maker decides to change the deployed model f at some point τ. However, up to τ, f was used to generate counterfactual recommendations. Will the recommendations still lead to the desired outcome under the competing model g? What are the expected additional costs due to the introduction of g? Will the cost depend on whether we use a *sparse* or *Data Support* counterfactual explanation machine?

(b) The decision maker is unsure about the correct classifier f and multiple models give almost identical hold-out test error. Will the explanations $E(\boldsymbol{x}; f)$ based on the model f generalize to a family of competing models g?

Point (b) refers to a concept usually known as 'researcher/practitioner degrees of freedom'. It describes that it is often not very clear why a certain classifier was chosen from a set of (potentially equally well performing) classifiers. On an individual end-user level, those choices make a difference and can have vast consequences. For example, they can determine whether someone gets a loan or not or whether a decision should be revised or not.

While there has been a sharp recent increase in the availability of methods that attempt to generate counterfactual recommendations (see section 2), there exists remarkably little work regarding their cost guarantees. At the same time, such guarantees might be a crucial element when deciding which method (*sparse* vs. *data support*) should be deployed in practice for consequential decisions with humans in the loop. This work attempts to close this gap.

Our contributions. In this paper, we challenge commonly held assumptions in the field of counterfactual explanations. We summarize our key contributions briefly:

- **Relating the costs of *Sparse* and *Data Support* counterfactuals.** We theoretically relate the cost of *sparse* and *Data Support* counterfactual recommendations. When the classifier is fixed, our result shows that *sparse* recommendations are provably less costly than those with *Data Support*. Under model multiplicity, we derive conditions which depend on the relative costs of both methods.
- **Cost guarantees for counterfactuals under model multiplicity**. We derive an upper bound on the cost of counterfactual explanations under model multiplicity. Our upper bound is stated in terms of the risk of both classifiers and most notably depends on how differently both classifiers assign negative predictions. Our result challenges the commonly held view that a counterfactual recommendations should have the *lowest possible cost* in general.
- **Empirical evaluation of the result.** Empirically, we compare *Data Support* and *Sparse* methods.

> Given one fixed classifier f, the *Data Support* recommendations are theoretically and empirically more costly, however, they are empirically more invariant to multiplicity of different models than *sparse* recommendations and result in semantically sensible recommendations.

Structure. In section 2 we give a categorization of different approaches. Section 3 contains theoretical cost guarantees, where we briefly discuss their implications. In section 4, we describe the compared models and evaluate them both quantitatively with respect to *cost* and *invariance to predictive multiplicity*, and qualitatively with respect to their semantics. Finally, section 5 concludes.

2 RELATED WORK

We denote the d-dimensional feature space as $\mathcal{X} = \mathbb{R}^d$ and the feature vector for observation i by $\boldsymbol{x} \in \mathcal{X}$ and the j-th dimesn denotes the The labels corresponding to the i-th observation are denoted by $y \in \mathcal{Y} = \{-1, +1\}$. Moreover, we assume two given pretrained, not identical, classifiers $f, g : \mathbb{R}^d \to \mathbb{R}$. Depending on the sign of $f(\boldsymbol{x})$ or the sign of $g(\boldsymbol{x})$ instances are classified. The goal is to find a counterfactual recommendation system for a given f, $E_f : \mathcal{X} \to \mathcal{X}$, generating counterfactuals $E(\boldsymbol{x}; f) = \tilde{\boldsymbol{x}}$, such that $\text{sign}(f(\boldsymbol{x})) \neq \text{sign}(f(E(\boldsymbol{x}; f)))$. We also introduce the following sets:

$$H_f^+ = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) > 0\}, H_f^- = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq 0\}$$
$$H_g^+ = \{\boldsymbol{x} \in \mathcal{X} : g(\boldsymbol{x}) > 0\}, H_g^- = \{\boldsymbol{x} \in \mathcal{X} : g(\boldsymbol{x}) \leq 0\}$$
$$D^+ = \{\boldsymbol{x} \in \mathcal{X} : y = +1\}, D^- = \{\boldsymbol{x} \in \mathcal{X} : y = -1\}$$

Exploring the local decision boundary. This line of work targets *purpose I* only. These approaches are based on perturbations and attempt to explain the sensitivity of a machine learning model to changes in its inputs by modelling the impact of local perturbations (Ribeiro et al., 2018; Adler et al., 2018; Fong and Vedaldi, 2017). Examples of perturbation-based approaches are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). However, none of them attempts to make recommendations for users who are directly affected by the classifications.

Counterfactual explanations. These works target both *purposes I and II*. Approaches dealing with tabular data rely on solving integer programming optimization problems (Ustun et al., 2019; Russell, 2019), use decision tree based classifiers (Tolomei et al., 2017), satisfiability modulo theory (Karimi et al., 2020a) or use data density approximation (via variational autoencoders) (Joshi et al., 2019; Pawelczyk et al., 2020). Other approaches ignore tabular data entirely (Grath et al., 2018; Laugel et al., 2017), but at least allow for conditionally immutable features (e.g. has a PhD) (Lash et al., 2017). To produce counterfactuals that take on reasonable values (e. g. non negative values for wage income) most approaches let decision makers specify the set of features and their respective support subject to change. We next aim to categorize most of the aforementioned approaches.

2.1 Sparse Approaches

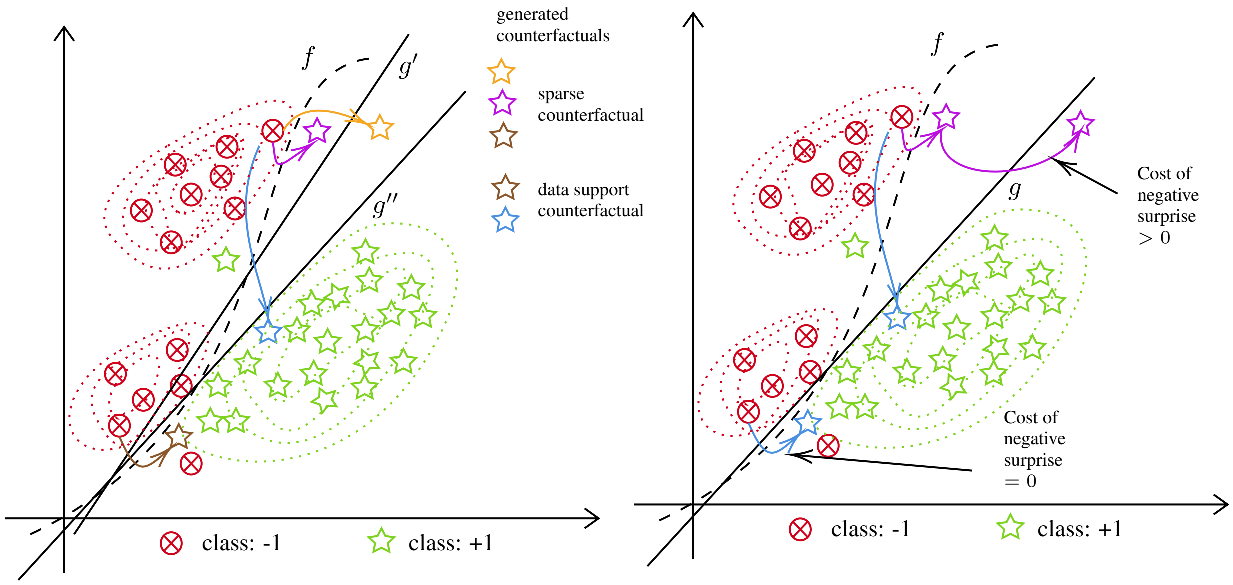
Definition 1. ***Sparse counterfactual recommendation.*** *Given inputs $\boldsymbol{x} \sim p_{data}$, a binary classifier $f(\boldsymbol{x})$ and a set of all possible counterfactual explanations, $\mathcal{E}_S = \{\tilde{\boldsymbol{x}} :\ sign(f(\tilde{\boldsymbol{x}})) = +1\}$, a sparse counterfactual recommendation is defined as $\boldsymbol{c}_S = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{E}_S} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_p$.*

Owing to interpretabiliy, p is usually 1 or 2. Several works have been put forth, relying on a variant of this definition (Laugel et al., 2017; Karimi et al., 2020a; Grath et al., 2018; Russell, 2019; Ustun et al., 2019; Wachter et al., 2017; Lash et al., 2017; Mothilal et al., 2020). In fact, a subset of these works additionally considered to restrict $\mathcal{E}_S$ further; for example some suggest to favour explanations over inputs that have shown to vary much in the past (Wachter et al., 2017), aim at generating diverse recommendations Mothilal et al. (2020) or allowed for having immutable inputs that could not be changed (e.g. *Gender*, *Age*) while searching for possible counterfactual recommendations (Lash et al., 2017; Ustun et al., 2019).

2.2 Data Support Approaches

Definition 2. ***(Data) Support counterfactual recommendation***. *Given inputs $\boldsymbol{x} \sim p_{data}$, a binary classifier $f(\boldsymbol{x})$ and a set of all admissible counterfactual explanations, $\mathcal{E}_D = \{\tilde{\boldsymbol{x}} :\ sign(f(\tilde{\boldsymbol{x}}) = +1 \text{ s.t. } p_{data}(\tilde{\boldsymbol{x}}) > 0\}$, a data supported counterfactual recommendation is defined as $\boldsymbol{c}_D = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{E}_D} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_p$.*

Definition 2 essentially demands that counterfactual recommendations should be supported by the true data distribution p_{data}. This is what is meant by *data support* and it comes at a cost since it is easy to see that $\boldsymbol{c}_D \geq \boldsymbol{c}_S$. In proposition 1 below we refine this statement. Additionally, consider figure 1 for an example. Of course, in practice we do not know p_{data} and therefore an explainability generator $E(\boldsymbol{x}; f)$ would need to take density estimation into account. Notice that this notion is also distinct from *actionability* (Ustun et al., 2019) or *plausibility* (Karimi et al., 2020a). They only demand that immutable inputs shall not be changed and that columns of $\tilde{\boldsymbol{x}}$ lie individually in a reasonable range. Per se, this does not imply $p_{data}(\tilde{\boldsymbol{x}}) > 0$. Laugel et al. (2019a) suggested density based evaluation measures to approximate


generated counterfactuals
sparse counterfactual
data support counterfactual
f
g'
g''
class: -1
class: +1
f
g
Cost of negative surprise > 0
Cost of negative surprise $= 0$
class: -1
class: +1

(a) *Sparse* vs. *Support* counterfactual recommendations.

(b) The *cost of negative surprise*.

Figure 1: (**a**) *Data support counterfactual* (def. 2) recommendations are more costly than *sparse counterfactual* recommendations (def. 1). See also proposition 1. (**b**) Suppose we made counterfactual recommendations under model f. If at some point τ we changed from f to g, then the *cost of negative surprise* is 0 for data support counterfactual recommendations while it is positive for sparse counterfactuals. Sparse counterfactual recommendations are more vulnerable to classifier uncertainty or classifier changes over time. Although more costly in the first place, *data support* counterfactuals are more transferable across different classifiers, i.e. they tend to have *lower cost of negative surprise.*

whether $p_{data}(\tilde{\boldsymbol{x}}) > 0$ holds. A small collection of works has devised methods to generate *data support counterfactual recommendations* (Joshi et al., 2019; Pawelczyk et al., 2020; Mahajan et al., 2019) using variational autoencoders (Kingma and Welling, 2013; Nazabal et al., 2018).

In light of the fact that counterfactual recommendation machines could have a huge impact on individuals' lives, there there exists remarkably little work regarding cost guarantees. With respect to our theoretical results the most relevant work is by Ustun et al. (2019). Proposition 2 is more general (see remark 1) since it considers the case of predictive multiplicity and nonlinear classifiers. In fact, it includes their result as a special case when the considered classifiers f and g coincide and are both linear. To the best of our knowledge, our work is the first that aims at relating the cost of *sparse* and *data support* counterfactuals.

3 COST GUARANTEES

3.1 Relation between *Sparse* and *Data support* Costs

Pawelczyk et al. (2020) establish empirically that there exists a trade-off between low-cost recommendations (*sparse*) and those with data support (they call them *attainable*). Here we give a theoretical underpinning of this empirical observation.

Suppose $\boldsymbol{x}$ is generated by a generative model h such that $h(\boldsymbol{z}) = \boldsymbol{x}$, where $\boldsymbol{z} \in \mathcal{Z} = \mathbb{R}^k$ are latent codes with $k < d$. As an example, $\boldsymbol{z}$ could be standard normal distributed. If the generative model was an autoencoder, this amounts to having a perfect encoder and decoder. As in Pawelczyk et al. (2020), we consider the following explanation mechanism to devise counterfactual recommendations via a nearest neighbour search in latent space:

$$f(h(\tilde{\boldsymbol{z}}))\text{ where }\tilde{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} \|h(z) - x\|. \qquad (1)$$

Then the following result holds.

Proposition 1 (Oracle cost inequality)**.** *The cost relation between* sparse *counterfactual recommendations and* data supported *counterfactual recommendations adheres:*

$$\boldsymbol{c}_D(\tilde{\boldsymbol{z}}) \leq 2 \cdot \boldsymbol{c}_S,$$

where *S* and *D* abbreviate *sparse* and *data support*, respectively. The proof can be found in appendix C. If we wish to obtain counterfactual recommendations with data support, proposition 1 suggests that there exists an extra cost, relative to the *sparse* counterfactual recommendations. In practice, however, a generative model is used for which the encoder and decoder parameters have to be estimated adequately. Therefore, the cost difference is likely to be higher since neither the encoder nor the decoder work perfectly. Our experiments in section 4 consolidate our findings.

3.2 Cost of Counterfactual Multiplicity

We start by stating the general objective. The goal is to find a minimal cost action c^* which alters the given classifiers' predicted labels from $\text{sign}((f(x)) = -1$ and $\text{sign}(g(x)) = -1$ to $+1$. More formally, we seek:

$$\begin{aligned} &c^*(f,g) = \underset{c\in\mathbb{R}^d}{\arg\min}\|c\| \text{ s.t.} \\ &\text{sign}(f(x+c)) = +1 \ \wedge \ \text{sign}(g(x+c)) = +1. \end{aligned} \tag{2}$$

Next, we state the main assumption used in proposition 1.

Assumption 1 (Pang (1997))**.** *There exist $\alpha > 0$ and $0 \le \gamma \le 1$ such that, for all $\boldsymbol{x}$,*

$$\begin{aligned} &dist(x, H_f^+ \cap H_g^+) \le \alpha \cdot max\{0, max(-f(x), -g(x))\}^\gamma, \\ &dist(x, H_f^+ \cap H_g^-) \le \alpha \cdot max\{0, max(-f(x), +g(x))\}^\gamma, \\ &dist(x, H_f^- \cap H_g^+) \le \alpha \cdot max\{0, max(+f(x), -g(x))\}^\gamma, \\ &dist(x, H_f^- \cap H_g^-) \le \alpha \cdot max\{0, max(+f(x), +g(x))\}^\gamma, \end{aligned}$$

where $dist(x, H) = \underset{s\in H}{min}\{\|\boldsymbol{x} - \boldsymbol{s}\|\}$.

The assumption states that a given point $\boldsymbol{x}$ is bounded by the so-called residual. We assume the residual provides a reasonable way to bound the distance from $\boldsymbol{x}$ to a point $\boldsymbol{s} \in H$ classified as $y = 1$ or $y = -1$. We proceed to define the quantity for which we give an upper bound.

Definition 3 (Cost of counterfactual multiplicity)**.** *The expected cost of counterfactual explanations under classifier multiplicity for classifiers $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is defined as,*

$$\overline{cost}(f,g)_{H_f^- \cup H_g^-} = \mathbb{E}_{H_f^- \cup H_g^-}[c^*(f,g)],$$

where the expectation is taken over the distribution of $\boldsymbol{x} \in H_f^- \cup H_g^-$. Analogs can be defined in which costs can be computed with respect to classifiers f or g only. For example, for the classifier f we would obtain $\overline{cost}(f)_{H_f^-} = \mathbb{E}_{H_f^-}[c^*(f)]$, where the objective in 2 would need to be altered appropriately. Definition 3 asks to find the expected minimum cost of counterfactual recommendations when we have to satisfy the constraint set out by two classifiers (see (2)).

Proposition 2 (Bounding costs of counterfactual multiplicity)**.** *Given assumption 1, the cost of counterfactual multiplicity under classifiers f and g, $\overline{cost}(f,g)_{H_f^- \cup H_g^-}$, is bounded from above such that,*

$$\begin{aligned} &\overline{cost}(f,g)_{H_f^- \cup H_g^-} \le \alpha \cdot 8^{1-\gamma} \\ &\cdot \Big[\underbrace{2 \cdot R_{H_f^-}(f) \cdot c^{max}_{H_f^-}(f) + 2 \cdot R_{H_g^-}(g) \cdot c^{max}_{H_g^-}(g)}_{\text{maximum risk of } f \text{ and } g} \\ &+ \underbrace{\pi_f \cdot c_{D^+}(f) + \pi_g \cdot c_{D^+}(g)}_{\text{false negative rates of } f \text{ and } g} \\ &- \underbrace{(1-\pi_f) \cdot c_{D^-}(f) - (1-\pi_g) \cdot c_{D^-}(g)}_{\text{true negative rates of } f \text{ and } g} \\ &+ \underbrace{\mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]}_{\substack{\boldsymbol{\Delta}(f,g)\text{: Discrepancy of } f \text{ and } g \\ \text{over neg. classified individuals}}}\Big]^\gamma, \textit{where} \end{aligned} \tag{3}$$

- $c_{D^+}(f) = \mathbb{E}_{H_f^- \cap D^+}[f(x)]$ *is the expected cost of counterfactual recommendations for individuals with $\boldsymbol{x} \in H_f^- \cap D^+$ (*false negative predictions*);*
- $c_{D^-}(f) = \mathbb{E}_{H_f^- \cap D^-}[f(x)]$ *is the expected cost of counterfactual recommendation for individuals with $\boldsymbol{x} \in H_f^- \cap D^-$ (*true negative predictions*);*
- $c^{max}_{H_f^-}(f) = \max_{x \in H_f^-}|f(x)|$ *denotes the max. cost of counterfactual recommendation for classifier f;*
- $\pi_f = Pr_{H_f^-}(y = 1)$ *is the false-omission rate of classifier f;*
- $R_{H_f^-}(f) = \pi_f Pr_{H^- \cap D^+}(f(x) \le 0) + (1 - \pi_f) Pr_{H_f^- \cap D^-}(f(x) > 0)$ *stands for the risk of classifier f for $x \in H_f^-$.*

Analogs can be defined for classifier g and the proof is given in appendix D. The expected cost of counterfactual explanations under predictive multiplicity does not directly depend on the overall classification error rate, but instead *focuses on those individuals for whom we made negative predictions.*

We would like to highlight the discrepancy term in proposition 2. Consider figure 2 for a more illustrative explanation of this term. Although the areas $H_f^+ \cap H_g^-$ and $H_f^- \cap H_g^+$ are already covered by the false and true negative rates of both classifiers, the discrepancy term counts them again. Intuitively, this is due to the fact that the red junctions in the upper left corner can neither be moved to $H_f^+ \cap H_g^-$ nor to $H_f^- \cap H_g^+$.

We would now like to take a step back and highlight some noteworthy real-world implications of this result:

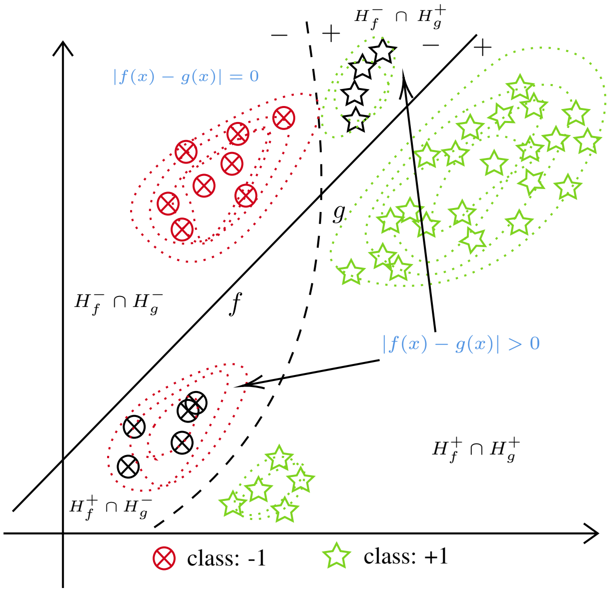


Figure 2: Discrepancy of f and g. For the area $H_f^- \cap H_g^-$ (upper left corner) the discrepancy $|f(x)-g(x)|$ is 0: both models agree that they should be classified negatively. The red junctions in the upper left corner can neither be moved to $H_f^+ \cap H_g^-$ nor to $H_f^- \cap H_g^+$. Since these areas are not admissible we have to pay an extra price for them to be moved to $H_f^+ \cap H_g^+$. This intuition is captured by the $\Delta(f, g)$ term in proposition 2.

- **Challenge minimal cost recommendations**. When assumption **A1** (Barocas et al. (2020); also mentioned in the introduction) is violated, our analysis suggests that finding sparse, minimal cost counterfactual recommendations with respect to a fixed classifier f fails to reflect the real expected cost of counterfactual recommendations.

- **Distort trust in automated ML**. If declined end users are initially issued a list of recommended feature changes and those changes turn out to be more costly due to, say, model updates over time, then this can severely distort trust in automated decision making systems. To the best of our knowledge, a legal framework for such cases does not exist, yet.

Remark 1. *If we take $\gamma = 1$ and f and g coincide, i.e. $f(x) = g(x)$ $\forall x$, then our result recovers theorem 3 in Ustun et al. (2019), where $\gamma = 1$ corresponds to the case where were we look at linear classifiers.*

Next we evaluate under which conditions any of the existing methods (*sparse* vs. *data support*) generate more robust counterfactual recommendations.

3.3 Relating the Cost of Negative Surprise for *Sparse* and *Data Support* Counterfactuals

We would like to find out what the additional cost induced by the classifier g would be. We call it the cost of *negative surprise*. Intuitively, it measures whether individuals subjected to a particular recommendation method (say *sparse* vs. *data support* recommendations) should be worried that their recommendation would change under a different classifier. If the classifiers g and f were to coincide for all instances $\boldsymbol{x}$, then the *cost of negative surprise* to all individuals would be 0 since no individual would need to exert additional effort/cost to satisfy a new constraint, which is illustrated in the right panel of figure 1.

Definition 4 (Inverse cost of negative surprise)**.** *The normalized inverse cost of negative surprise under model multiplicity for classifiers $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ under method $M = \{D, S\}$ is defined as:*

$$\overline{s}(f, g)_M = \left[\frac{\mathbb{E}_{H_f^- \cup H_g^-}[c^*(f(x), g(x))]_M}{\mathbb{E}_{H_f^-}[c^*(f(x))]_M} \right]^{-1} \in (0, 1].$$

The inverse cost is a measure of invariance of a counterfactual recommendation to different classifiers and ideally evaluates to 1. This happens when $\mathbb{E}_{H_f^- \cup H_g^-}[c^*(f, g)]_M = \mathbb{E}_{H_f^-}[c^*(f)]_M$, that is, in expectation there will be no additional changes to the cost of counterfactual recommendations due to the introduction of a competing classifier g.

Remark 2. *Definition 4 appears cumbersome, however, it allows us to use a lower bound for $\mathbb{E}_{H_f^- \cup H_g^-}[c^*(f, g)]$, which depends on $\mathbb{E}_{H_f^-}[c^*(f)]$ and $\mathbb{E}_{H_g^-}[c^*(g)]$.*

Proposition 3 (Negative surprise for sparse and data support recommendations)**.** *For simplicity of the statement, suppose $\gamma = 1$. If $\mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_S = \mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_D$ and*

$$\frac{\mathbb{E}_{H_g^-}[c^*(g)]_D}{\mathbb{E}_{H_f^-}[c^*(f)]_D} < \frac{\mathbb{E}_{H_g^-}[c^*(g)]_S}{\mathbb{E}_{H_f^-}[c^*(f)]_S}.$$

then we must have that:

$$1 \geq \overline{s}(f, g)_S > \overline{s}(f, g)_D.$$

The proof of Proposition 3 can be found in appendix E. It suggests a direct way to check whether counterfactual suggestions generated by the *sparse* methods are less prone to negative surprise than those generated by the *Data Support* camp. Note that the *sparse* explanation method could trivially satisfy this condition by generating high cost recommendations for the classifier g. For example, a *sparse* method could push the red junctions in the upper left of figure 2 all the way to the bottom right of the plot. However, by definition 1 *sparse* explanations mechanisms are not set out to do so. It would indeed be counterproductive for the generation of counterfactual recommendations that end-users can realistically translate into lived realities.

This result implies an interesting question (for future research): *Can we generate invariant counterfactual recommendations with minimal costs*? In the following, we do not suggest a new way to do so, but we evaluate whether existing methods already do.

4 EXPERIMENTS

Data sets. We conduct extensive quantitative and qualitative evaluations on two different realistic classification settings: (i) 'Give Me Some Credit' and (ii) HELOC.

"Give Me Some Credit". This data set contains 10 inputs and 8 are related to the individual's financial history. We assume that the inputs are mutable and their types are *count* and *positive continuous*, respectively. The remaining 2 features, *age* and *# dependencies*, are immutable.

HELOC. This data set has 23 inputs of which all of them are related to the individual's financial history. All the inputs are treated as count variables. We treat the *ExternalRiskEstimate*, *MSinceOldestTradeOpen* and *AverageMInFile* as immutable since they are not under the individual's direct control. The remaining inputs are treated as *mutable*. The data set originally holds 10000 observations, but after dropping observations with missing instances we are left with $n = 8291$.

Methods. We choose three methods and compare across three dimensions. First, what is the associated cost of the generated counterfactual recommendations. Second, what is the individual cost of negative surprise, i.e. how well do the generated counterfactual explanations generalize to other models? Third, do the generated recommendations make semantically sense?

(*Sparse methods*) The first chosen method is classifier agnostic and conducts a greedy nearest-neighbour search. It chooses the closest counterfactual recommendations measured by the ℓ_2-norm (Laugel et al., 2017) (GS). The second method was suggested by Ustun et al. (2019) (AR). They use integer programming tools subject to cost function (4) and (5). While their method is restricted to linear classifiers, it also works for tabular data. (*Data Support methods*) The last method is classifier agnostic and was concurrently suggested by Joshi et al. (2019); Pawelczyk et al. (2020); Mahajan et al. (2019). We use the method as suggested in Pawelczyk et al. (2020) (OURS).They use a special type of variational autoencoder (VAE) (Kingma and Welling, 2013; Nazabal et al., 2018) for counterfactual search. The idea is to train a (V)AE that deals well with tabular data and to leverage the latent space representation to search for counterfactual recommendations.

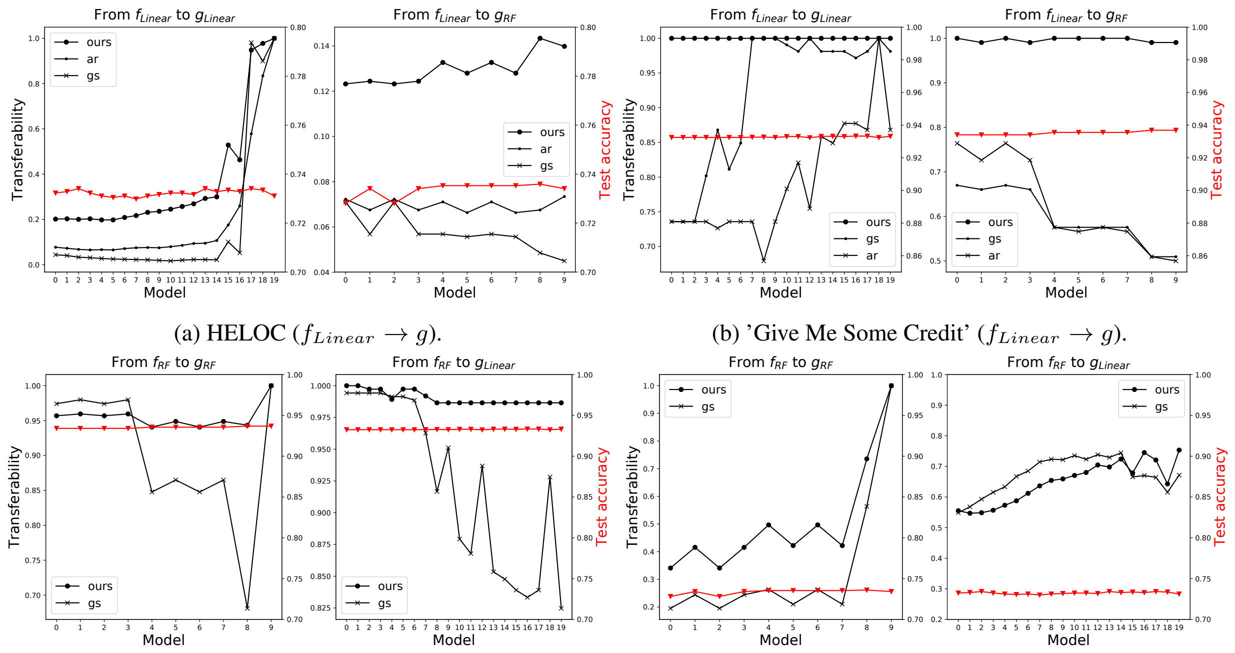
4.1 The (Local) Cost of Negative Surprise

In this section, we investigate different models' ability to generate counterfactual recommendations that generalize well across different classifiers. If they generalize well, then they have a low cost of negative surprise. To do so, we distinguish the following two cases: (a) Holding the hypothesis class fixed, will the initially generated counterfactual recommendation under the model f_{θ_1} generalize to changes in the parameters θ while the risk of both classifiers stays approximately the same, i.e. $R(f_{\theta_1}) \approx R(f_{\theta_2})$? (b) Will the initially generated counterfactual recommendation under the hypothesis class $\mathcal{F}$ (e.g. regularized linear models) generalize to a model g from a different hypothesis class (e.g. random forest) while $R(f) \approx R(g)$? For the experiment described in (a), we do not transfer the counterfactual recommendation to any model, but only to those from the ϵ-level set. For the experiments in (b) we extend the below definition to models outside the hypothesis class $\mathcal{F}$.

Definition 5 (ϵ-level set (Marx et al., 2019))**.** *Given any classifier f and a hypothesis class $\mathcal{F}$, the ϵ-level set around f is the set of all models $g \in \mathcal{F}$ that make at most $\hat{R}(f) + \epsilon$ mistakes over the training data.*

We choose $\epsilon = +/{-0.05}$. To generate the models from the ϵ-level set we use the `cv grid search` method from `scikit learn`. We then use models g within the set and check whether the counterfactual recommendations generated based on f are equally valid under g. Particularly, we check two different hypothesis classes: $\mathcal{F}_{Linear}$ and $\mathcal{F}_{RandomForest} = \mathcal{F}_{RF}$. In practice we generate $\tilde{x}(f) := E(x; f)$ and compute $\mathcal{T} = 1/n_E \cdot \mathbb{I}[f(\tilde{x}(f)) = g(\tilde{x}(f))]$, where n_E is the number of individuals for which counterfactual recommendations are computed and $\mathbb{I}(\cdot)$ denotes the indicator function.

Next, the results are shown in figure 3. The left y-axis depicts how many counterfactual explanations were transferable from model f to model g (It measures $\mathcal{T}$.). The x-axis indicates the model number, and the right y-axis (red graph) shows the model's corresponding test accuracy. The models are usually ordered (in the left column) so that the rightmost model corresponds to the model we used to generate the counterfactual recommendations in



(a) HELOC ($f_{Linear} \to g$).

(b) 'Give Me Some Credit' ($f_{Linear} \to g$).

(c) 'Give Me Some Credit' ($f_{RF} \to g$).

(d) HELOC ($f_{RF} \to g$).

Figure 3: **Invariance to predictive multiplicity**. We generate counterfactual explanations according to model f and then check whether they are still valid under model g. Left axis: percentage of counterfactual explanations that are robust to model changes. Right axis: model accuracy on hold-out test set. Generally we observe that counterfactuals from OURS are more invariant to model changes than those from GS and AS.

the first place. To summarize the results, we would like to stress several points.

(a) **Importance of invariant explanations**. This exercise underlines the importance to learn counterfactual recommendations that are invariant to small (within hypothesis class) and large (between hypothesis class) model perturbations. This is important since the effect of predictive multiplicity makes the model with respect to which we generate counterfactual recommendations look almost arbitrary.

(b) **Data supported counterfactuals are more often model invariant**. The OURS model generates the most robust recommendations: it outperforms GS and AR on all tasks and almost all classifiers. It performs a little worse on the HELOC data set when transferring from $\mathcal{F}_{RF}$ to $\mathcal{F}_{Linear}$ (right panel in figure 3d).

In this section, we have empirically investigated whether counterfactual recommendations generalize across models and are thus robust to predictive multiplicity. We found that the data support based methods had superior generalization capabilities. Proposition 1 and 2 suggest that these recommendations should also be more costly. We investigate this shortly in the following section.

4.2 Costs of Counterfactual Recommendations

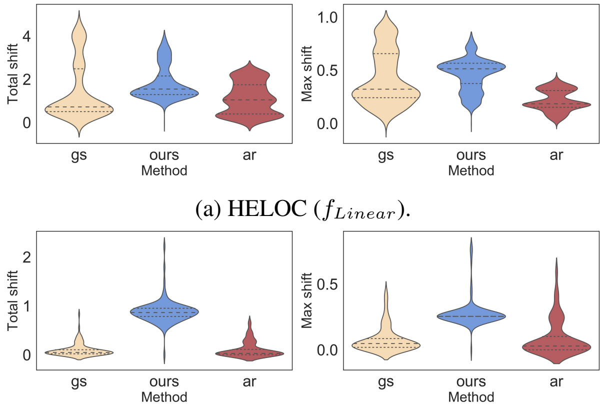
In order to evaluate the cost of counterfactual suggestions across different models, we use the following two measures (Pawelczyk et al., 2020):

$$cost_1(\tilde{x}; x) = \sum_j |(Q_j(\tilde{x}_j) - Q_j(x_j)|, \qquad (4)$$

$$cost_2(\tilde{x}; x) = \underset{j}{max}\ |Q_j(\tilde{x}_j) - Q_j(x_j)|, \qquad (5)$$

where the subscript denotes the j-th component of x. The total percentile shift in (4) can be thought of as a baseline measure for how attainable a certain counterfactual suggestion might be. The maximum percentile shift (MS) in (5) across all free features reflects the maximum difficulty across all inputs that are subject to change.

Figure 4 shows the resulting plots. The left panel shows violinplots for the distribution of total percentile shifts and the right panel shows these plots for the maximum percentile shift. From the plots it becomes clear that the OURS method generates counterfactual recommendations that tend to have both higher total and maximum percentile shifts. This holds for both data sets.



(a) HELOC (f_{Linear}).

(b) ’Give Me Some Credit’ (f_{Linear}).

Figure 4: **Costs of counterfactual recommendations**. The data density based method OURS (it uses a VAE) generates more costly counterfactual recommendations than GS and AR.

In the next section, we briefly investigate why OURS works better in producing invariant recommendations. We will also understand why it generates higher costs.

4.3 Why Data Supported Counterfactuals tend to Generate more Invariant Explanations

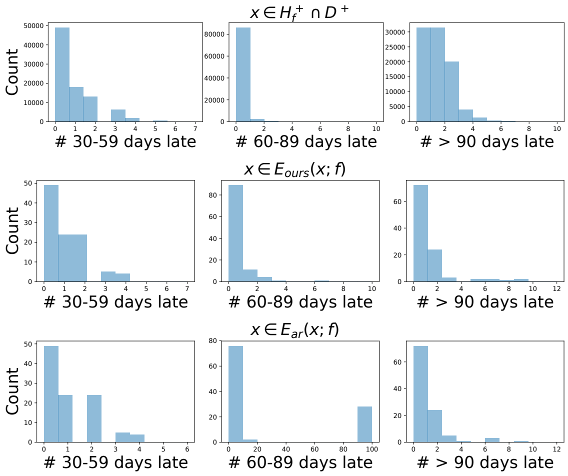


Figure 5: **Timeliness and counterfactual suggestions** for the ”Give Me Some Credit” data. (Top row) Histogram of three inputs related to timeliness of loan payments for individuals from $H_f^+ \cap D^+$. (Middle & Bottom row) Histogram for counterfactual recommendations on the test set for explanations from OURS (middle) and AR (bottom).

On robustness. Next, we zoom into one particular set of inputs. We look at the distribution of three inputs: `30-59 days late`, `60-89 days late`, `>89 days late`, which one could summarize as *timeliness* of individuals’ payments. The first row in figure 5 shows the distribution of the three inputs for which we have that $H_f^+ \cap D^+$: in words, correctly classified individuals make their loan payments on time. The counterfactual recommendations generated by OURS, $E_{ours}(\boldsymbol{x}; f)$, second row of figure 5, follow this distribution quite closely. The last row shows the distribution induced by $E_{ar}(\boldsymbol{x}; f)$, which is not close to the one of the correctly classified individuals for `# 60-89 days late`.

On costs. Recall from section 4.2 and figure 4 that OURS’ recommendations are more costly. Now, let us consider figure 5 again. To generate lower cost recommendations, AR needs to be very close to the original inputs. Thus, it often suggests to leave the timeliness inputs unchanged (third row in figure 5) or only change a subset of them. This, however, appears counter intuitive (we discuss this issue further in appendix A).

5 CONCLUSION

In light of the fact that counterfactual recommendations can have an huge impact on individuals’ lives, there existed remarkably little work regarding cost guarantees for existing methods. In this work, we have taken a step towards filling this void. We theoretically analyzed the cost of counterfactual recommendations for *sparse* and *data supported* counterfactual recommendations. Most notably, we obtained the following insights: first, *data supported* counterfactual recommendations are at least as costly as *sparse* ones. Second, if assumption **A1** (classifier is stable) is violated, the cost of counterfactual recommendations under model multiplicity can be substantially higher than under one fixed model. Therefore, counterfactual recommendations are ideally based on (explanation) models that causally (and thus invariantly) relate inputs to targets to avoid the impact of predictive multiplicity on counterfactuals.

Our results have thus guided us to an interesting question for future research: *can one generate invariant counterfactual recommendations with minimal costs*?

To establish trustworthy (semi-) automated ML systems with humans in the loop, it is crucial to provide counterfactual recommendations with cost guarantees, which humans can rely on when working towards their goals. Therefore, we hope that our work can help practitioners make more informed decisions on which type of recommendation method to choose in the future.

References

Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *FAT**, 2020.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *CVPR*, pages 3429–3437, 2017.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *NeurIPS workshop: Challenges and Opportunities for AI in Fin. Services*, 2018.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *AISTATS*, 2020a.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. *arXiv preprint arXiv:2002.06278*, 2020b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013.

Michael T Lash, Qihang Lin, Nick Street, Jennifer G Robinson, and Jeffrey Ohlmann. Generalized inverse classification. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 162–170. SIAM, 2017.

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

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Issues with post-hoc counterfactual explanations: a discussion. *ICML: Workshop on Human in the Loop Learning*, 2019a.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *IJCAI*, 2019b.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

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

Charles T Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. *arXiv preprint arXiv:1909.06677*, 2019.

Peter McCullagh and John A. Nelder. *Generalized linear models*. CRC Press, 1989.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, 2020.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*, 2018.

Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79:299–332, 1997.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning counterfactual explanations for tabular data. In *WWW*. ACM, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *SIGKDD*. ACM, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018.

Christopher Russell. Efficient search for diverse coherent explanations. In *FAT**. ACM, 2019.

Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *SIGKDD*. ACM, 2017.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *FAT**. ACM, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):2018, 2017.