Decisions, Counterfactual Explanations and Strategic Behavior

Stratis Tsirtsis and Manuel Gomez-Rodriguez

Max Planck Institute for Software Systems
{stsirtsis, manuelgr}@mpi-sws.org

Abstract

As data-driven predictive models are increasingly used to inform decisions, it has been argued that decision makers should provide explanations that help individuals understand what would have to change for these decisions to be beneficial ones. However, there has been little discussion on the possibility that individuals may use the above *counterfactual explanations* to invest effort strategically and maximize their chances of receiving a beneficial decision. In this paper, our goal is to find policies and counterfactual explanations that are optimal in terms of utility in such a strategic setting. We first show that, given a pre-defined policy, the problem of finding the optimal set of counterfactual explanations is NP-hard. Then, we show that the corresponding objective is nondecreasing and satisfies submodularity and this allows a standard greedy algorithm to enjoy approximation guarantees. In addition, we further show that the problem of jointly finding both the optimal policy and set of counterfactual explanations reduces to maximizing a non-monotone submodular function. As a result, we can use a recent randomized algorithm to solve the problem, which also offers approximation guarantees. Finally, we demonstrate that, by incorporating a matroid constraint into the problem formulation, we can increase the diversity of the optimal set of counterfactual explanations and incentivize individuals across the whole spectrum of the population to self improve. Experiments on synthetic and real lending and credit card data illustrate our theoretical findings and show that the counterfactual explanations and decision policies found by our algorithms achieve higher utility than several competitive baselines.

1 Introduction

Whenever a bank decides to offer a loan to a customer, a judge decides to grant bail to a person, or a company decides to hire a new employee, the decision is increasingly informed by a data-driven predictive model. In all these high-stakes applications, the goal of the decision maker is to take decisions that maximize a given utility function while the goal of the predictive model is to provide accurate predictions of the outcomes from a set of observable features. For example, a bank may decide whether or not to offer a loan to a customer using the model's estimate of the probability that the customer would repay the loan.

In this context, there has been a tremendous excitement on the potential of data-driven predictive models to enhance decision making in high-stakes applications. However, there has also been a heated debate about their lack of transparency and explainability (Doshi-Velez and Kim, 2017; Weller, 2017; Lipton, 2018; Gunning and Aha, 2019; Rudin, 2019). As a result, there already exists a legal requirement to grant individuals who are subject to (semi)-automated decision making the *right-to-explanation* in the European Union (Voigt and Von dem Bussche, 2017; Wachter et al., 2017a). With this motivation, there has been a flurry of work on interpretable machine learning (Ribeiro et al., 2016; Koh and Liang, 2017; Lundberg and Lee, 2017; Chakraborty et al., 2017; Murdoch et al., 2019; Wachter et al., 2017b; Karimi et al., 2019; Mothilal et al., 2020), which has predominantly focused on developing methods to find explanations for the predictions made by a predictive model. Within this line of work, the work most closely related to ours (Wachter et al., 2017b; Karimi et al., 2019; Mothilal et al., 2020) aims to find counterfactual explanations that help individuals understand what would have to change for a predictive model to make a positive prediction about them. In our work, rather than explaining predictions, we pursue the development of methods to find counterfactual explanations for the decisions taken by a decision maker[1], which are ultimately what individuals who are subject to (semi)-automated decision making typically care about. In this context, we will assume that the decision maker takes decisions based on low dimensional feature vectors since, in many realistic scenarios, the data is summarized by just a small number of summary statistics (*e.g.*, FICO scores) (Hardt et al., 2016b; Liu et al., 2018).

Once we focus on explaining decisions, we cannot overlook the possibility that individuals may use these explanations to invest effort strategically in order to maximize their chances of receiving a beneficial decision. However, this is also an opportunity for us to find counterfactual explanations that help individuals to self-improve and eventually increase the utility of a decision policy, as noted by several studies in economics (Coate and Loury, 1993; Fryer and Loury, 2013; Hu and Chen, 2018) and, more recently, in the computer science literature (Kleinberg and Raghavan, 2019; Perdomo et al., 2020; Tabibian et al., 2020). For example, if a bank explains to a customer that, if she reduces her credit card debt by 20%, she will receive the loan she is applying for, she may feel compelled to reduce her overall credit card debt by the proposed percentage to pay less interest, improving her financial situation, and this will eventually increase the profit the bank makes when she is able to successfully return the loan. This is in contrast with previous work on interpretable machine learning, which have ignored the influence that (counterfactual) explanations (of predictions by a predictive model) may have on the accuracy of predictive models and the utility of the decision policies.

Our contributions. We cast the above problem as a Stackelberg game in which the decision maker moves first and shares her counterfactual explanations before individuals best-respond to these explanations and invest effort to receive a beneficial decision. Under this problem formulation, we first show that, given a pre-defined policy, the problem of finding the optimal set of counterfactual explanations is NP-hard by using a novel reduction of the Set Cover problem (Karp, 1972). Then, we show that the corresponding objective function is monotone and submodular and, as a direct consequence, it readily follows that a standard greedy algorithm offers approximation guarantees. In addition, we show that, given a pre-defined set of counterfactual explanations, the optimal policy is deterministic and can be computed in polynomial time. Moreover, building on this result, we can reduce the problem of jointly finding both the optimal policy and set of counterfactual explanations to maximizing a non-monotone submodular function. As a consequence, we can use a recent randomized algorithm to solve the problem, which also offers approximation guarantees. Further, we demonstrate that, by incorporating a matroid constraint into the problem formulation, we can increase the diversity of the optimal set of counterfactual explanations and incentivize individuals across the whole spectrum of the population to self improve. Experiments using real lending and credit card data illustrate our theoretical findings and show that the counterfactual explanations and decision policies found by the above algorithms achieve higher utility than several competitive baselines[2].

2 Problem Formulation

Given an individual with a feature vector $\boldsymbol{x} \in \{1, ..., n\}^d$ and a (*ground-truth*) label $y \in \{0, 1\}$, we assume a decision $d(\boldsymbol{x}) \in \{0, 1\}$ controls whether the corresponding label is *realized*[3]. This setting fits a variety of real-world scenarios, where continuous features are often discretized into (percentile) ranges. For example, in university admissions, the decision specifies whether a student is admitted ($d(\boldsymbol{x}) = 1$) or rejected ($d(\boldsymbol{x}) = 0$); the label indicates whether the student completes the program ($y = 1$) or drops out ($y = 0$) upon acceptance; and the feature vector ($\boldsymbol{x}$) may include her GRE scores, undergraduate GPA percentile, or research experience. Throughout the paper, we will denote the set of feature values as $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, where $m = n^d$ denotes the number of feature values, and assume that the number of features d is small, as discussed previously.

Each decision is sampled from a decision policy $d(\boldsymbol{x}) \sim \pi(d \,|\, \boldsymbol{x})$, where, for brevity, we will write $\pi(\boldsymbol{x}) = \pi(d = 1 \,|\, \boldsymbol{x})$. For each individual, the label y is sampled from a conditional probability distribution $y \sim P(y \,|\, \boldsymbol{x})$ and, without loss of generality, we index the feature values in decreasing order with respect to their corresponding outcome, *i.e.*, $i < j \Rightarrow P(y = 1 \,|\, \boldsymbol{x}_i) \geq P(y = 1 \,|\, \boldsymbol{x}_j)$. Moreover, we adopt a

[1]These counterfactual explanations help individuals understand what would have to change in order to receive a beneficial decision, rather than a positive prediction.

[2]Data and code to reproduce the results in our paper are publicly available at `https://github.com/Networks-Learning/strategic-decisions`.

[3]Without loss of generality, we assume each feature takes n different values.

Stackelberg game-theoretic formulation in which each individual with initial feature value $\boldsymbol{x}_i$ receives a (counterfactual) explanation from the decision maker by means of a feature value $\mathcal{E}(\boldsymbol{x}_i) \in \mathcal{A} \subseteq \mathcal{P}_\pi := \{\boldsymbol{x} \in \mathcal{X} : \pi(\boldsymbol{x}) = 1\}$ before she (best-)responds[4]. This formulation fits a variety of real-world applications. For example, insurance companies often provide online car insurance simulators that, on the basis of a customer's initial feature value $\boldsymbol{x}_i$, let the customer know whether they are eligible for a particular deal. In case the customer does not qualify, the simulator could provide a counterfactual example $\mathcal{E}(\boldsymbol{x}_i)$ under which the individual is guaranteed to be eligible. In the remainder, we will refer to $\mathcal{A}$ as the set of counterfactual explanations and, for each individual with initial feature value $\boldsymbol{x}_i$, we will assume she does not know anything about the other counterfactual explanations $\mathcal{A} \backslash \mathcal{E}(\boldsymbol{x}_i)$ other individuals may receive nor the decision policy $\pi(\boldsymbol{x})$.

Now, let $c(\boldsymbol{x}, \mathcal{E}(\boldsymbol{x}_i))$ be the cost[5] an individual pays for changing from $\boldsymbol{x}_i$ to $\mathcal{E}(\boldsymbol{x}_i)$ and $b(\pi, \boldsymbol{x}) = \mathbb{E}_{d \sim \pi(d \,|\, x)}[d(\boldsymbol{x})]$ be the (immediate) benefit she obtains from a policy π, which is just the probability that the individual receives a positive decision. Then, following Tabibian et al. (2020), each individual's best response is to change from her initial feature value $\boldsymbol{x}_i$ to $\mathcal{E}(\boldsymbol{x}_i)$ iff the gained benefit she would obtain outweighs the cost she would pay for changing features, *i.e.*,

$$\mathcal{E}(\boldsymbol{x}_i) \in \{\boldsymbol{x}_j \in \mathcal{X} \,:\, b(\pi, \boldsymbol{x}_j) - c(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq b(\pi, \boldsymbol{x}_i)\} := \mathcal{R}(\boldsymbol{x}_i),$$

and it is to keep her initial feature value $\boldsymbol{x}_i$ otherwise. Here, we will refer to $\mathcal{R}(\boldsymbol{x}_i)$ as the *region of adaptation*. Then, at a population level, the above best response results into a transportation of mass between the original feature distribution $P(\boldsymbol{x})$ and a new feature distribution $P(\boldsymbol{x} \,|\, \pi, \mathcal{A})$ induced by the policy π and the counterfactual explanations $\mathcal{A}$. More specifically, we can readily derive an analytical expression for the induced feature distribution in terms of the original feature distribution, *i.e.*, for all $\boldsymbol{x}_j \in \mathcal{X}$,

$$P(\boldsymbol{x}_j \,|\, \pi, \mathcal{A}) = P(\boldsymbol{x}_j)\mathbb{I}(\mathcal{R}(\boldsymbol{x}_j) \cap \mathcal{A} = \emptyset) + \sum_{i \in [m]} P(\boldsymbol{x}_i)\mathbb{I}(\mathcal{E}(\boldsymbol{x}_i) = \boldsymbol{x}_j \wedge \boldsymbol{x}_j \in \mathcal{R}(\boldsymbol{x}_i)),$$

Similarly as in previous work (Corbett-Davies et al., 2017; Valera et al., 2018; Kilbertus et al., 2019; Tabibian et al., 2020), we will assume that the decision maker is rational, has access to (an estimation of) the original feature distribution $P(\boldsymbol{x})$, and aims to maximize the (immediate) utility $u(\pi, \gamma)$, which is the expected overall profit she obtains, *i.e.*,

$$u(\pi, \mathcal{A}) = \mathbb{E}_{\boldsymbol{x} \sim P(\boldsymbol{x} \,|\, \pi, \mathcal{A}), y \sim P(y \,|\, \boldsymbol{x}), d \sim \pi(\boldsymbol{x})}\left[y d(\boldsymbol{x}) - \gamma d(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x} \sim P(\boldsymbol{x} \,|\, \pi, \mathcal{A})}\left[\pi(\boldsymbol{x})(P(y = 1 \,|\, \boldsymbol{x}) - \gamma)\right], \quad (1)$$

where $\gamma \in (0, 1)$ is a given constant reflecting economic considerations of the decision maker. For example, in university admissions, the term $\pi(\boldsymbol{x})P(y = 1 \,|\, \boldsymbol{x})$ is proportional to the expected number of students who are admitted and complete the program, the term $\pi(\boldsymbol{x})\gamma$ is proportional to the number of students who are admitted, and γ measures the cost of education in units of graduated students. As a direct consequence, given a feature value $\boldsymbol{x}_i$ and a set of counterfactual explanations $\mathcal{A}$, we can conclude that, if $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} \neq \emptyset$, the decision maker will decide to provide the counterfactual explanation $\mathcal{E}(\boldsymbol{x}_i)$ that provides the largest utility gain under the assumption that individuals best respond, *i.e.*,

$$\mathcal{E}(\boldsymbol{x}_i) = \underset{\boldsymbol{x} \in \mathcal{A} \cap \mathcal{R}(\boldsymbol{x}_i)}{\operatorname{argmax}} \; P(y \,|\, \boldsymbol{x}) \text{ for all } \boldsymbol{x}_i \in \mathcal{X} \setminus \mathcal{P}_\pi \text{ such that } \mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} \neq \emptyset, \quad (2)$$

and, if $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} = \emptyset$, we arbitrarily assume that $\mathcal{E}(\boldsymbol{x}_i) = \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{A}} c(\boldsymbol{x}_i, \boldsymbol{x})$[6].

Given the above preliminaries, our goal is to help the decision maker to first find the optimal set of counterfactual explanations $\mathcal{A}$ for a pre-defined policy in Section 3 and then both the optimal policy π and set of counterfactual explanations $\mathcal{A}$ in Section 4.

Remarks. Given an individual with initial feature value $\boldsymbol{x}$, one may think that, by providing the counterfactual explanation $\mathcal{E}(\boldsymbol{x}) \in \mathcal{A} \cap \mathcal{R}(\boldsymbol{x})$ that gives the largest utility gain, the decision maker is not acting in the individual's best interest but rather selfishly. This is because there may exist another

[4]In practice, individuals with initial feature values $\boldsymbol{x}_i$ such that $\pi(\boldsymbol{x}) = 1$ may not receive any explanation since they are guaranteed to receive a positive decision.

[5]In practice, the cost for each pair of feature values may be given by a parameterized function.

[6]Note that, if $\mathcal{A} \cap \mathcal{R}(\boldsymbol{x}_i) = \emptyset$, the individual's best response is to keep her initial feature value $\boldsymbol{x}_i$ and thus any choice of counterfactual explanation $\mathcal{E}(\boldsymbol{x}_i)$ leads to the same utility.

counterfactual explanation $\boldsymbol{x}' \in \mathcal{A} \cap \mathcal{R}(\boldsymbol{x})$ with lower cost for the individual, *i.e.*, $c(\boldsymbol{x}, \boldsymbol{x}') \leq c(\boldsymbol{x}, \mathcal{E}(\boldsymbol{x}))$. However, in our work, we argue that the provided counterfactual explanations help the individual to achieve a greater self-improvement and this is likely to result in a superior long-term well-being. For example, if a bank explains to a customer that she will receive the loan she is applying for if she reduces her credit card debt by 10%, rather than 20%, even though the corresponding feature values are both within the region of adaptation $\mathcal{R}(\boldsymbol{x})$ of her original features $\boldsymbol{x}$, the customer will be more likely to default and this will very negatively impact her long-term well-being.

As argued very recently (Miller et al., 2019; Tabibian et al., 2020), due to Goodhart's law, the conditional probability $P(y \,|\, \boldsymbol{x})$ may change after individuals (best)-respond if the features $\boldsymbol{x}$ are noncausal. Moreover, Miller et al. (2019) have argued that (best-)responses to noncausal and causal features correspond to gaming and improvement, respectively. In this work, for simplicity, we assume that $P(y \,|\, \boldsymbol{x})$ does not change, however, it would be very interesting to lift this assumption in future work.

3 Finding the optimal counterfactual explanations for a policy

In this section, our goal is to find the optimal set of counterfactual explanations $\mathcal{A}^*$ for a pre-defined policy π, *i.e.*,

$$\mathcal{A}^* = \underset{\mathcal{A} \subseteq \mathcal{P}_\pi \,:\, |\mathcal{A}| \leq k}{\operatorname{argmax}} \; u(\pi, \mathcal{A}), \tag{3}$$

where the cardinality constraint on the set of counterfactual explanations balances the decision maker's obligation to be transparent with trade secrets (Barocas et al., 2020). More specifically, note that, without this constraint, an adversary could reverse-engineer the entire decision policy $\pi(\boldsymbol{x})$ by impersonating individuals with different feature values $\boldsymbol{x}$ (cs-).

As it will become clearer in the experimental evaluation in Section 6, our results may persuade decision makers to be transparent about their decision policies, something they are typically reluctant to be despite the increasing legal requirements, since we show that transparency increases the utility of the policies. Moreover, throughout this section, we will assume that the decision maker who picks the pre-defined policy is rational[7] and the policy is outcome monotonic[8][9] (Tabibian et al., 2020). Outcome monotonicity just implies that, the higher an individual's outcome $P(y = 1 \,|\, \boldsymbol{x})$, the higher their chances of receiving a positive decision $\pi(\boldsymbol{x})$.

Unfortunately, using a novel reduction of the Set Cover problem (Karp, 1972), the following theorem reveals that we cannot expect to find the optimal set of counterfactual explanations in polynomial time (proven in Appendix B.1):

Theorem 1 *The problem of finding the optimal set of counterfactual explanations that maximizes utility under a cardinality constraint is NP-Hard.*

Even though Theorem 1 is a negative result, we will now show that the objective function in Eq. 3 satisfies a set of desirable properties, *i.e.*, non-negativity, monotonicity and submodularity[10], which allow a standard greedy algorithm to enjoy approximation guarantees at solving the problem. To this aim, with a slight abuse of notation, we first express the objective function as a set function $f(\mathcal{A}) = u(\pi, \mathcal{A})$, which takes values over the ground set of counterfactual explanations, $\mathcal{P}_\pi$. Then, we have the following proposition (proven in Appendix B.2):

Proposition 2 *The function f is non-negative, submodular and monotone.*

The above result directly implies that the standard greedy algorithm (Nemhauser et al., 1978) (refer to Algorithm 1 in Appendix C) will find a solution $\mathcal{A}$ to the problem such that $f(\mathcal{A}) \geq (1 - 1/e) f(\mathcal{A}^*)$, where $\mathcal{A}^*$ is the optimal set of counterfactual explanations. Moreover, since the greedy algorithm computes the marginal difference of f for at most m elements per iteration and, following from the proof of Proposition 2, the marginal difference $f(\mathcal{A} \cup \{\boldsymbol{x}\}) - f(\mathcal{A})$ can be computed in $\mathcal{O}(m)$, then it immediately follows that, in our problem, the greedy algorithm has an overall complexity of $\mathcal{O}(km^2)$.

[7]Note that, if the decision maker is rational and her goal is to maximize the utility, as defined in Eq. 1, then, for all $\boldsymbol{x} \in \mathcal{X}$ such that $P(y = 1 \,|\, \boldsymbol{x}) < \gamma$, it holds that $\pi(\boldsymbol{x}) = 0$.

[8]A policy π is called outcome monotonic if $P(y = 1 \,|\, \boldsymbol{x}_i) \geq P(y = 1 \,|\, \boldsymbol{x}_j) \Leftrightarrow \pi(\boldsymbol{x}_i) \geq \pi(\boldsymbol{x}_j) \; \forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$.

[9]If the policy π is deterministic, our results also hold for non outcome monotonic policies.

[10]A function $f : 2^{\mathcal{X}} \to \mathbb{R}$ is submodular if for every $\mathcal{A}, \mathcal{B} \subseteq \mathcal{X} : \mathcal{A} \subseteq \mathcal{B}$ and $x \in \mathcal{X} \setminus \mathcal{B}$ it holds that $f(\mathcal{A} \cup \{x\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{x\}) - f(\mathcal{B})$.

4 Finding the optimal policy and counterfactual explanations

In this section, our goal is to jointly find the optimal decision policy and set of counterfactual explanations $\mathcal{A}^*$, *i.e.*,

$$\pi^*, \mathcal{A}^* = \underset{(\pi, \mathcal{A}): \mathcal{A} \subseteq \mathcal{P}_\pi \wedge |\mathcal{A}| \leq k}{\operatorname{argmax}} \; u(\pi, \mathcal{A}) \tag{4}$$

where, similarly as in the previous section, k is the maximum number of counterfactual explanations the decision maker is willing to provide to the population to balance the right to explanation with trade secrets. By jointly optimizing both the decision policy and the counterfactual explanations, we may obtain an additional gain in terms of utility in comparison with just optimizing for the set of counterfactual explanations given the optimal decision policy in a non-strategic setting, as shown in Figure 5 in Appendix D. Moreover, as we will show in the experimental evaluation in Section 6, this additional gain will be significant.

Similarly as in Section 3, we cannot expect to find the optimal policy and set of counterfactual explanations in polynomial time. More specifically, we have the following negative result, which easily follows from Proposition 4 and slightly extending the proof of Theorem 1:

Theorem 3 *The problem of jointly finding both the optimal policy and the set of counterfactual explanations that maximize utility under a cardinality constraint is NP-hard.*

However, while the problem of finding both the policy and the set of counterfactual explanations appears significantly more challenging than the problem of finding just the set of counterfactual explanations given a pre-defined policy (refer to Eq. 3), the following proposition shows that the problem is not inherently *harder*. More specifically, for each possible set of counterfactual explanations, it shows that the policy that maximizes the utility can be easily computed (proven in Appendix B.3):

Proposition 4 *Given a set of counterfactual explanations $\mathcal{A} \subseteq \mathcal{Y} := \{\boldsymbol{x} \in \mathcal{X} : P(y \,|\, \boldsymbol{x}) \geq \gamma\}$[11], the policy $\pi^*_{\mathcal{A}} = \operatorname{argmax}_{\pi: \mathcal{A} \subseteq \mathcal{P}_\pi} u(\pi, \mathcal{A})$ that maximizes the utility is deterministic and can be found in polynomial time,* i.e.,

$$\pi^*_{\mathcal{A}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in \mathcal{A} \vee \{\boldsymbol{x}' \in \mathcal{A} : P(y = 1 \,|\, \boldsymbol{x}') > P(y = 1 \,|\, \boldsymbol{x}) \wedge c(\boldsymbol{x}, \boldsymbol{x}') \leq 1\} = \emptyset\} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The above result implies that, to set all the values of the optimal decision policy, we only need to perform $\mathcal{O}(km)$ comparisons. Moreover, it reveals that, in contrast with the non strategic setting, the optimal policy given a set of counterfactual explanations is not a deterministic threshold rule with a single threshold (Corbett-Davies et al., 2017; Valera et al., 2018), *i.e.*,

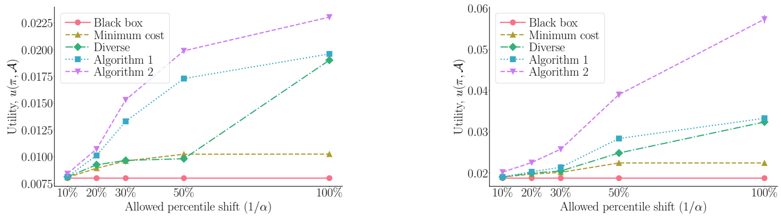
$$\pi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } P(y = 1 \,|\, \boldsymbol{x}) \geq \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

but rather a more conservative deterministic decision policy that does not depend only on the outcome $P(y = 1 \,|\, \boldsymbol{x})$ and γ but also on the cost individuals pay to change features. Moreover, we can build up on the above result to prove that the problem of finding the optimal decision policy and set of counterfactual explanations can be reduced to maximizing a non-monotone submodular function. To this aim, let $\pi^*_{\mathcal{A}}$ be the optimal policy induced by a given set of counterfactual explanations $\mathcal{A}$, as in Proposition 4, and define the set function $h(\mathcal{A}) = u(\pi^*_{\mathcal{A}}, \mathcal{A})$ over the ground set $\mathcal{Y}$. Then, we have the following proposition (proven in Appendix B.4):

Proposition 5 *The function h is non-negative, submodular and non-monotone.*

Fortunately, there exist efficient algorithms with global approximation guarantees for maximizing a non-monotone submodular function under cardinality constraints. For example, Buchbinder et al. (2014) have proposed a randomized polynomial time algorithm (refer to Algorithm 2 in Appendix C) that can find a solution $\mathcal{A}$ such that $h(\mathcal{A}) \geq (1/e) h(\mathcal{A}^*)$, where $\mathcal{A}^*$ and $\pi^*_{\mathcal{A}^*}$ are the optimal set of counterfactual explanations and decision policy, respectively. Moreover, since the above randomized algorithm has a complexity of $\mathcal{O}(km)$ and, following from the proof of Proposition 5, the marginal difference of h can be computed in $\mathcal{O}(m)$, it readily follows that, in our problem, the algorithm has a complexity of $\mathcal{O}(km^2)$.

[11] Since the decision maker is rational, she will never provide an explanation that contributes negatively to her utility.



(a) Lending dataset (b) Credit dataset

Figure 1: Utility achieved by five types of decision policies and counterfactual explanations against the value of the parameter α, which controls how difficult it is to change feature values, in the lending and credit datasets. In panel (a), the number of feature values is $m = 400$ and, in panel (b), it is $m = 3200$. In both panels, we set $k = 0.05m$ and we repeat each experiment 20 times.

5 Increasing the diversity of the counterfactual explanations

In many cases, decision makers may like to ensure that individuals across the whole spectrum of the population are incentivized to self-improve. For example, in a loan scenario, the bank may use age group as a feature to estimate the probability that a customer repays the loan, however, it may like to deploy a decision policy that incentivizes individuals across all age groups in order to improve the financial situation of all. To this aim, the decision maker can increase the diversity of the optimal set of counterfactual explanations by incorporating a matroid constraint into the problem formulation, rather than a cardinality constraint.

Formally, consider disjoint sets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_l$ such that $\bigcup_i \mathcal{X}_i = \mathcal{X}$ and integers $d_1, d_2, \ldots, d_l$ such that $k = \sum_i d_i$. Then, a partition matroid is the collection of sets $\{S \subseteq 2^{\mathcal{X}} : |S \cap \mathcal{X}_i| \leq d_i \; \forall i \in [l]\}$. In the loan example, the decision maker could search for a set of counterfactual explanations $\mathcal{A}$ within a partition matroid where each one of the $\mathcal{X}_i$'s corresponds to the feature values covered by each age group and $d_i = k/l \; \forall i \in [l]$. This way, the set of counterfactual explanations $\mathcal{A}$ would include explanations for every age group.

In this case, the decision maker could rely on a variety of polynomial time algorithms with global guarantees for submodular function maximization under matroid constraints, *e.g.*, the algorithm by Calinescu et al. (2011).

6 Experiments

In this section, we evaluate Algorithms 1 and 2 using real loan and credit card data and show that the counterfactual explanations and decision policies found by our algorithms achieve higher utility than several competitive baselines. Appendix E contains additional experiments on synthetic data.

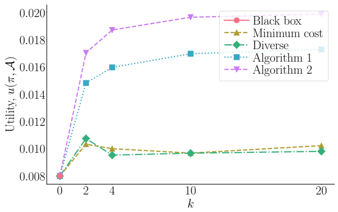
Experimental setup. We experiment with two publicly available datasets: (i) the *lending* dataset (len), which contains information about all accepted loan applications in LendingClub during the 2007-2018 period and (ii) the *credit* dataset (Yeh and Lien, 2009), which contains information about a bank's credit card payoffs[12]. For each accepted loan applicant (or credit card holder), we use various demographic information and financial status indicators as features $\boldsymbol{x}$ and the current loan status (or credit payoff status) as label y. Appendix F contains more details on the specific features we used in each dataset and also describes the procedure we followed to approximate $P(y \,|\, \boldsymbol{x})$ and estimate the cost function $c(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

In our experiments, we compare the utility of the following decision policies and counterfactual explanations:

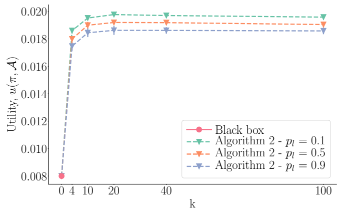
— *Black box:* decisions are taken by the optimal decision policy in the non-strategic setting, given by Eq. 6, and individuals do not receive any counterfactual explanations.

— *Minimum cost:* decisions are taken by the optimal decision policy in the non-strategic setting, given by Eq. 6, and individuals receive counterfactual explanations of minimum cost with respect to their initial feature values, similarly as in previous work (Ustun et al., 2019; Tolomei et al., 2017; Karimi et al., 2019). More specifically, we cast the problem of finding the set of counterfactual explanations

[12] We used a version of the credit dataset preprocessed by Ustun et al. (2019)



(a) Utility vs. k



(b) Utility vs. k under leakage

Figure 2: Number of counterfactual explanations and information leakage. Panel (a) shows the utility achieved by five types of decision policies and counterfactual explanations against the number of counterfactual explanations k. Panel (b) shows the utility achieved by Algorithm 2 against the number of counterfactual explanations k for several values of the leakage probability p_l. In both panels, we use the lending dataset, the number of feature values is $m = 400$, we set $\alpha = 2$, we repeat each experiment involving randomization 20 times.

as the minimization of the weighted average cost individuals pay to change their feature values to the closest counterfactual explanation, *i.e.*,

$$\mathcal{A}_{mc} = \underset{\mathcal{A} \subseteq \mathcal{P}_\pi \,:\, |\mathcal{A}| \leq k}{\operatorname{argmin}} \sum_{\boldsymbol{x}_i \in \mathcal{X} \setminus \mathcal{P}_\pi} P(\boldsymbol{x}_i) \min_{\boldsymbol{x}_j \in \mathcal{A}} c(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

and realize that this problem is a version of the k-median problem, which we can solve using a greedy heuristic (Solis-Oba, 2006).

— *Diverse:* decisions are taken by the optimal decision policy in the non-strategic setting, given by Eq. 6, and individuals receive a set of diverse counterfactual explanations of minimum cost with respect to their initial feature values, similarly as in previous work (Russell, 2019; Mothilal et al., 2020), *i.e.*,

$$\mathcal{A}_{d} = \underset{\mathcal{A} \subseteq \mathcal{P}_\pi \,:\, |\mathcal{A}| \leq k}{\operatorname{argmax}} \sum_{\boldsymbol{x}_i \in \mathcal{X} \setminus \mathcal{P}_\pi} P(\boldsymbol{x}_i) \mathbb{I}(\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} \neq \emptyset),$$

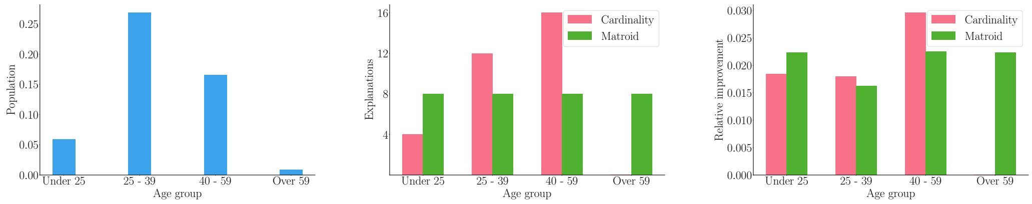
To solve the above problem, we realize it can be reduced to the weighted version of the maximum coverage problem, which can be solved using a well-known greedy approximation algorithm (Hochbaum and Pathria, 1998).

— *Algorithm 1:* decisions are taken by the optimal decision policy in the non-strategic setting, given by Eq. 6, and individuals receive counterfactual explanations given by Eq. 2, where $\mathcal{A}$ is found using Algorithm 1.

— *Algorithm 2:* decisions are taken by the decision policy given by Eq. 5 and individuals receive counterfactual explanations given by Eq. 2, where $\mathcal{A}$ is found using Algorithm 2.

Results. We start by comparing the utility achieved by each of the decision policies and counterfactual explanations for several values of the parameter α in both datasets. Figure 1 summarizes the results, which show that Algorithm 1 and Algorithm 2 consistently outperform all baselines and, as the cost of adapting to feature values with higher outcome values decreases (smaller α), the competitive advantage by jointly optimizing the decision policy and the counterfactual explanations (Algorithm 2) grows significantly. This competitive advantage is more apparent in the credit card dataset because it contains non actionable features (*e.g.*, credit overdue counts) and, under the optimal decision policy in the non-strategic setting, it is difficult to incentivize individuals who receive a negative decision to improve by just optimizing the set of counterfactual explanations they receive. Appendix F.3 elaborates more on this insight by visualizing the transportation of mass induced by the policies and counterfactual explanations used in Algorithm 1 and Algorithm 2 and Appendix F.4 provides specific examples of counterfactual explanations provided by Algorithm 1 and the minimum cost baseline.

Next, we focus on the lending dataset and evaluate the sensitivity of our algorithms. First, we measure the influence that the number of counterfactual explanations has on the utility achieved by each of the decision policies and counterfactual explanations. As shown in Figure 2(a), our algorithms just need a small number of counterfactual explanations to provide significant gains in terms of utility with respect to all the baselines. Second, we challenge the assumption that individuals do not share the counterfactual



(a) Rejected population by age (b) Explanations by age (c) Relative improvement by age

Figure 3: Increasing the diversity of the provided counterfactual explanations. Panel (a) shows the population per age group, rejected by the optimal threshold policy in the non strategic setting. Panel (b) shows a comparison of the age distribution of counterfactual explanations in $\mathcal{A}$ produced by the greedy algorithm under a cardinality and a matroid constraint. Panel (c) shows the relative improvement of each age group. In all panels, we use the credit dataset and we set $k = 32$ and $\alpha = 2$.

explanations they receive with other individuals with different feature values. To this end, we assume that, given the set of counterfactual explanations $\mathcal{A}$ found by Algorithm 2, individuals with initial feature value $\boldsymbol{x}$ receive the counterfactual explanation $\mathcal{E}(\boldsymbol{x}) \in \mathcal{A}$ given by Eq. 2 and, with probability p_l, they also receive an additional explanation $\mathcal{E}'(\boldsymbol{x})$ picked at random from $\mathcal{A}$ and they follow the counterfactual explanation that benefits them the most. Figure 2(b) summarizes the results for several values of p_l and number of counterfactual explanations, which show that the policies and explanations provided by Algorithm 2 present a significant utility advantage even when the leakage probability p_l is large.

Finally, we focus on the credit dataset and consider a scenario in which a bank aims not only to continue providing credit to the customers that are more likely to repay but also provide explanations that incentivize individuals across all age groups to maintain their credit. To this end, we incorporate a partition matroid constraint that ensures the counterfactual explanations are diverse across age groups, as described in Section 5, and use a slightly modified version of Algorithm 1 to solve the constrained problem (Nemhauser et al., 1978), which enjoys a 1/2 approximation guarantee. Figure 3 summarizes the results, which show that: (i) optimizing under a cardinality constraint leads to an unbalanced set of explanations, favoring the more populated age groups (25 to 59) while completely ignoring the recourse potential of individuals older than 60; (ii) the relative group improvement, defined as $\sum_{\boldsymbol{x}_i \in \mathcal{X}_z \setminus \mathcal{P}_\pi} P(\boldsymbol{x}_i)[P(y \,|\, \boldsymbol{x}_j^i) - P(y \,|\, \boldsymbol{x}_i)] / \sum_{\boldsymbol{x}_i \in \mathcal{X}_z \setminus \mathcal{P}_\pi} P(\boldsymbol{x}_i)$, where $\mathcal{X}_z$ is the set of feature values corresponding to age group z and $\boldsymbol{x}_j^i$ is the best response of individuals with initial feature value $\boldsymbol{x}_i \in \mathcal{X}_z$, is more balanced across age groups, showing that the matroid constraint can be used to generate counterfactual explanations that help the entire spectrum of the population to self-improve.

7 Conclusions

In this paper, we have designed several algorithms that allow us to find the decision policies and counterfactual explanations that maximize utility in a setting in which individuals who are subject to the decisions taken by the policies use the counterfactual explanations they receive to invest effort strategically. Moreover, we have experimented with synthetic and real lending and credit card data and shown that the counterfactual explanations and decision policies found by our algorithms achieve higher utility than several competitive baselines.

By uncovering a previously unexplored connection between strategic machine learning and interpretable machine learning, our work opens up many interesting directions for future work. For example, we have adopted a specific type of mechanism to provide counterfactual explanations (*i.e.*, one feature value per individual using a Stackelberg formulation). A natural next step would be to extend our analysis to other types of mechanisms fitting a variety of real-world applications. Moreover, we have assumed that the cost individuals pay to change features is given. However, our algorithms would be more effective if we develop a methodology to reliably estimate the cost function from real observational (or interventional) data. In our work, we have assumed that features take discrete values and individuals who are subject to the decisions do not share information between them. It would be interesting to lift these assumptions, extend our analysis to real-valued feature values, and develop decision policies and counterfactual explanations that are robust to information sharing between individuals (refer to Figure 2(c)). Finally, by assuming that $P(y \,|\, \boldsymbol{x})$ does not change after individuals best respond, we are implicitly assuming that the features $\boldsymbol{x}$ are causal. However, in practice, this assumption is likely to be violated, as recently noted by Miller et al. (2019). It would be worth exploring the use of counterfactual explanations to distinguish between noncausal and causal features.

References

Credit score simulator. https://www.creditkarma.com/tools/credit-score-simulator/.

Lending club dataset. https://www.kaggle.com/wordsforthewise/lending-club/version/3.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555, 2011.

Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1433–1452. SIAM, 2014.

Gruia Calinescu, Chandra Chekuri, Martin Pal, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE, 2017.

S. Coate and G. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1993.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

R. Fryer and G. Loury. Valuing diversity. *Journal of Political Economy*, 2013.

David Gunning and David W Aha. Darpa's explainable artificial intelligence program. *AI Magazine*, 40 (2):44–58, 2019.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016b.

Dorit S Hochbaum and Anu Pathria. Analysis of the greedy approach in problems of maximum k-coverage. *Naval Research Logistics (NRL)*, 45(6):615–627, 1998.

L. Hu and Y. Chen. A short-term intervention for long-term fairness in the labor market. In *WWW*, 2018.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

Richard M Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer, 1972.

Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *AISTATS*, 2019.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Advances in neural information processing systems*, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 2017.

John Miller, Smitha Milli, and Moritz Hardt. Strategic adaptation to classifiers: A causal perspective. *arXiv preprint arXiv:1910.10362*, 2019.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical programming*, 14(1):265–294, 1978.

Juan C Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *arXiv preprint arXiv:2002.06673*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

Roberto Solis-Oba. Approximation algorithms for the k-median problem. In *Efficient Approximation and Online Algorithms*, pages 292–320. Springer, 2006.

Behzad Tabibian, Stratis Tsirtsis, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *Arxiv:1905.09239*, 2020.

Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 465–474, 2017.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*, pages 1769–1778, 2018.

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

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017a.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017b.

Adrian Weller. Challenges for transparency. 2017.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

A Further related work

Our work builds upon previous work on interpretable machine learning, strategic machine learning, and machine-assisted decision making.

Most previous work on interpretable machine learning has focused on one of the two following types of explanations: feature-based explanations (Ribeiro et al., 2016; Koh and Liang, 2017; Lundberg and Lee, 2017) or counterfactual explanations (Wachter et al., 2017b; Karimi et al., 2019; Mothilal et al., 2020). Feature-based explanations help individuals understand the importance each feature has on a particular prediction, typically through local approximation, while counterfactual explanations help them understand what features would have to change for a predictive model to make a positive prediction about them. While there is not yet an agreement on what constitutes a *good* post-hoc explanation in the literature on interpretable machine learning, counterfactual explanations are gaining prominence because they place no constraints on the model complexity, do not require model disclosure, facilitate actionable recourse, and seem to automate compliance with the law (Barocas et al., 2020). Motivated by these desirable properties, our work focuses on counterfactual explanations and sheds light on the possibility of using explanations to increase the utility of a decision policy, uncovering a previously unexplored connection between interpretable machine learning and the nascent field of strategic machine learning.

Similarly as in our work, previous work on strategic machine learning also assumes that individuals may use knowledge, gained by transparency, to invest effort strategically in order to receive either a positive prediction (Brückner and Scheffer, 2011; Dalvi et al., 2004; Dong et al., 2018; Hardt et al., 2016a; Hu et al., 2019; Milli et al., 2019; Miller et al., 2019; Perdomo et al., 2020) or a beneficial decision (Kleinberg and Raghavan, 2019; Tabibian et al., 2020). However, none of this previous work focuses on finding (counterfactual) explanations and they assume full transparency—individuals who are subject to (semi)-automated decision making can observe the entire predictive model or the decision policy. As a result, their formulation is fundamentally different and their technical contributions are orthogonal to ours.

In the machine-assisted decision making literature, the distinction between decisions and predictions has not been made explicit until very recently (Corbett-Davies et al., 2017; Kilbertus et al., 2019; Kleinberg et al., 2018; Mitchell et al., 2018; Tabibian et al., 2020; Valera et al., 2018). However, previous work has focused on the design of optimal decision policies rather than (counterfactual) explanations.

B Proofs

B.1 Proof of Theorem 1

Consider an instance of the Set Cover problem with a set of elements $\mathcal{U} = \{u_1, \ldots, u_n\}$ and a collection $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\} \subseteq 2^{\mathcal{U}}$ such that $\bigcup_{i\in[m]} \mathcal{S}_i = \mathcal{U}$. In the decision version of the problem, given a constant k, we need to answer the question whether there are at most k sets from the collection $\mathcal{S}$ such that their union is equal to $\mathcal{U}$ or not. With the following procedure, we show that any instance of that problem can be transformed to an instance of the problem of finding the optimal set of counterfactual explanations, defined in Eq. 3, in polynomial time.

Consider $n + m$ feature values corresponding to the n elements of $\mathcal{U}$ and the m sets of $\mathcal{S}$. Moreover, denote the first n feature values as $\boldsymbol{x}_{u_1}, \ldots, \boldsymbol{x}_{u_n}$ and the remaining m as $\boldsymbol{x}_{\mathcal{S}_1}, \ldots, \boldsymbol{x}_{\mathcal{S}_m}$. We set the decision maker's parameter γ to some positive constant less than 1. Then, we set the outcome probabilities $P(y = 1|\boldsymbol{x}_{u_i}) = \gamma\ \forall i \in [n]$ and $P(y = 1|\boldsymbol{x}_{\mathcal{S}_i}) = 1\ \forall i \in [m]$ and the policy values $\pi(\boldsymbol{x}_{u_i}) = 0\ \forall i \in [n]$ and $\pi(\boldsymbol{x}_{\mathcal{S}_i}) = 1\ \forall i \in [m]$. This way, the portion of utility the decision-maker obtains from the first n feature values is zero, while the portion of utility she obtains from the remaining m is proportional to $1 - \gamma$. Regarding the cost function, we set $c(\boldsymbol{x}_{u_i}, \boldsymbol{x}_{\mathcal{S}_j}) = 0\ \forall (\boldsymbol{x}_{u_i}, \boldsymbol{x}_{\mathcal{S}_j}) : u_i \in S_j$, $c(\boldsymbol{x}_{u_i}, \boldsymbol{x}_{u_i}) = 0\ \forall i \in [n]$, and all the remaining values of the cost function to 2. Finally, we set the initial feature value distribution to $P(\boldsymbol{x}_{u_i}) = \frac{1}{n}\ \forall i \in [n]$ and $P(\boldsymbol{x}_{\mathcal{S}_i}) = 0\ \forall i \in [m]$. A toy example of this transformation is presented in Figure 4.

In this setting, it easy to observe that an individual with initial feature value $\boldsymbol{x}_{u_i}$ is always rejected at first and has the ability to move to a new feature value $\boldsymbol{x}_{\mathcal{S}_j}$ recommended to her iff $c(\boldsymbol{x}_{u_i}, \boldsymbol{x}_{\mathcal{S}_j}) \leq 1 \Leftrightarrow u_i \in \mathcal{S}_j$. Also, we can easily see that the transformation of instances can be done in $\mathcal{O}((m + n)^2)$ time.

Now, assume there exists an algorithm that optimally solves the problem of finding the optimal set of counterfactual explanations in polynomial time. Given the aforementioned instance and a maximum number of counterfactual explanations k, the utility $u(\pi, \mathcal{A})$ achieved by the set of counterfactual explanations $\mathcal{A}$ the algorithm returns can fall into one of the following two cases:

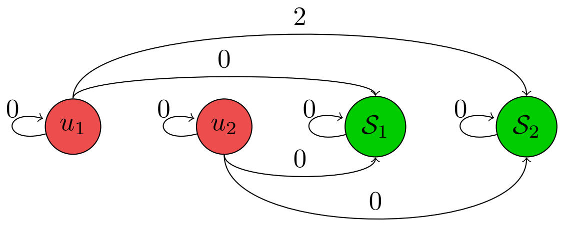


Figure 4: Consider that $\mathcal{U} = \{u_1, u_2\}$ and $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2\}$ with $\mathcal{S}_1 = \{u_1, u_2\}$, $\mathcal{S}_2 = \{u_2\}$. The red feature values have initial population $P(\boldsymbol{x}) = 1/2$, $\pi(\boldsymbol{x}) = 0$ and $P(y = 1 \,|\, \boldsymbol{x}) = \gamma$ while for the green feature values it is $P(\boldsymbol{x}) = 0$, $\pi(\boldsymbol{x}) = 1$ and $P(y = 1 \,|\, \boldsymbol{x}) = 1$. The edges represent the cost between feature values corresponding to sets and their respective elements while all the non-visible pairwise costs are equal to 2.

1. $u(\pi, \mathcal{A}) = 1 - \gamma$. This can happen only if all individuals, according to the induced distribution $P(\boldsymbol{x} \,|\, \pi, \mathcal{A})$, have moved to some of the feature values $\boldsymbol{x}_{\mathcal{S}_j}$, *i.e.*, for all $\boldsymbol{x}_{u_i}$ with $i \in [n]$, there exists $\boldsymbol{x}_{\mathcal{S}_j}$ with $j \in [m]$ such that $\boldsymbol{x}_{\mathcal{S}_j} \in \mathcal{A} \wedge c(\boldsymbol{x}_{u_i}, \boldsymbol{x}_{\mathcal{S}_j}) \leq 1$ with $|\mathcal{A}| \leq k$. As a consequence, if we define $\mathcal{S}' = \{\mathcal{S}_j : \boldsymbol{x}_{\mathcal{S}_j} \in \mathcal{A}\}$, it holds that for all u_i with $i \in [n]$, there exists $\mathcal{S}_j$ with $j \in [m]$ such that $\mathcal{S}_j \in \mathcal{S}' \wedge u_i \in \mathcal{S}_j$ and therefore $\mathcal{S}'$ is a set cover with $|\mathcal{S}'| = |\mathcal{A}| \leq k$.
2. $u(\pi, \mathcal{A}) < 1 - \gamma$. This can happen only if every possible set of k counterfactual explanations leaves the individuals of at least one feature value $\boldsymbol{x}_{u_i}$ with a best-response of not following the counterfactual explanation they were given, *i.e.*, for all $\mathcal{A} \subseteq \mathcal{P}_\pi$ such that $|\mathcal{A}| \leq k$, there exists $\boldsymbol{x}_{u_i}$ with $i \in [n]$ such that, for all $\boldsymbol{x}_{\mathcal{S}_j} \in \mathcal{A}$, it holds that $c(\boldsymbol{x}_{u_i}, \boldsymbol{x}_{\mathcal{S}_j}) > 1$. Equivalently, it holds that for all $\mathcal{S}' \subseteq \mathcal{S}$ such that $|\mathcal{S}'| \leq k$, there exists u_i with $i \in [n]$ such that for all $\mathcal{S}_j \in \mathcal{S}'$, it holds that $u_i \notin \mathcal{S}_j$ and therefore there does not exist a set cover of size less or equal than k.

The above directly implies that we can have a decision about any instance of the Set Cover problem in polynomial time, which is a contradiction unless $P = NP$. This concludes the reduction and proves that the problem of finding the optimal set of counterfactual explanations for a given policy is NP-Hard.

B.2 Proof of Proposition 2

It readily follows that the function f is non-negative from the fact that, if the decision maker is rational, it holds that $\pi(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$ such that $P(y = 1 \,|\, \boldsymbol{x}) < \gamma$.

Now, consider two sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}_\pi : \mathcal{A} \subseteq \mathcal{B}$ and a feature value $\boldsymbol{x} \in \mathcal{P}_\pi \setminus \mathcal{B}$. Also, let $\mathcal{E}_\mathcal{S}(\boldsymbol{x}_i)$ be the counterfactual explanation given to the individuals with initial feature value $\boldsymbol{x}_i$ under a set of counterfactual explanations $\mathcal{S}$. It is easy to see that the marginal difference $f(\mathcal{S} \cup \{\boldsymbol{x}\}) - f(\mathcal{S})$ can only be affected by individuals with initial features $\boldsymbol{x}_i$ such that $\boldsymbol{x}_i \notin \mathcal{P}_\pi$, $\boldsymbol{x} \in \mathcal{R}(\boldsymbol{x}_i)$ and $\boldsymbol{x} = \mathcal{E}_{\mathcal{S} \cup \{\boldsymbol{x}\}}(\boldsymbol{x}_i)$. Moreover, we can divide all of these individuals into two cases:

1. $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} = \emptyset$: in this case, the addition of $\boldsymbol{x}$ to $\mathcal{A}$ causes a change in their best-response from $\boldsymbol{x}_i$ to $\boldsymbol{x}$ contributing to the marginal difference of f by a factor $P(\boldsymbol{x}_i)[P(y = 1 \,|\, \boldsymbol{x}) - \gamma - \pi(\boldsymbol{x}_i)(P(y = 1 \,|\, \boldsymbol{x}_i) - \gamma)]$. However, considering the marginal difference of f under the set of counterfactual explanations $\mathcal{B}$, three subcases are possible:
   (a) $\mathcal{E}_\mathcal{B}(\boldsymbol{x}_i) \in \mathcal{R}(\boldsymbol{x}_i) \wedge P(y = 1 \,|\, \mathcal{E}_\mathcal{B}(\boldsymbol{x}_i)) > P(y = 1 \,|\, \boldsymbol{x})$: the contribution to the marginal difference of f is zero.
   (b) $\mathcal{E}_\mathcal{B}(\boldsymbol{x}_i) \in \mathcal{R}(\boldsymbol{x}_i) \wedge P(y = 1 \,|\, \mathcal{E}_\mathcal{B}(\boldsymbol{x}_i)) \leq P(y = 1 \,|\, \boldsymbol{x})$: the contribution to the marginal difference of f is $P(\boldsymbol{x}_i)[P(y = 1 \,|\, \boldsymbol{x}) - P(y = 1 \,|\, \mathcal{E}_\mathcal{B}(\boldsymbol{x}_i))]$. Since π is outcome monotonic, $\mathcal{E}_\mathcal{B}(\boldsymbol{x}_i) \in \mathcal{P}_\pi$ and $\boldsymbol{x}_i \notin \mathcal{P}_\pi$, it holds that

$$
\begin{aligned}
&P(y = 1 \,|\, \mathcal{E}_\mathcal{B}(\boldsymbol{x}_i)) \geq P(y = 1 \,|\, \boldsymbol{x}_i) \Rightarrow \\
&P(y = 1 \,|\, \mathcal{E}_\mathcal{B}(\boldsymbol{x}_i)) - \gamma \geq P(y = 1 \,|\, \boldsymbol{x}_i) - \gamma > \pi(\boldsymbol{x}_i)[P(y = 1 \,|\, \boldsymbol{x}_i) - \gamma].
\end{aligned}
$$

Therefore, it readily follows that

$$P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i))] < \\ P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - \gamma - \pi(\boldsymbol{x}_i)(P(y=1\,|\,\boldsymbol{x}_i) - \gamma)].$$

(c) $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{B} = \emptyset$: the contribution to the marginal difference of f is $P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - \gamma - \pi(\boldsymbol{x}_i)(P(y=1\,|\,\boldsymbol{x}_i) - \gamma)]$.

2. $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} \neq \emptyset \wedge P(y=1\,|\,\boldsymbol{x}) > P(y=1\,|\,\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i))$: In this case, the addition of $\boldsymbol{x}$ to $\mathcal{A}$ causes a change in their best-response from $\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i)$ to $\boldsymbol{x}$ contributing to the marginal difference of f by a factor $P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - P(y=1\,|\,\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i))]$. Considering the marginal difference of f under the set of counterfactual explanations $\mathcal{B}$, two subcases are possible:

   (a) $\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i) \in \mathcal{R}(\boldsymbol{x}_i) \wedge P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i)) > P(y=1\,|\,\boldsymbol{x})$: the contribution to the marginal difference of f is zero.

   (b) $\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i) \in \mathcal{R}(\boldsymbol{x}_i) \wedge P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i)) \leq P(y=1\,|\,\boldsymbol{x})$. Then, the contribution of those individuals to the marginal difference of f is $P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i))]$. Since $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{R}(\boldsymbol{x}_i) \cap \mathcal{A} \neq \emptyset$, it readily follows that

$$P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i)) \geq P(y=1\,|\,\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i)) \Rightarrow \\ P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - P(y=1\,|\,\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i))] \geq \\ P(\boldsymbol{x}_i)[P(y=1\,|\,\boldsymbol{x}) - P(y=1\,|\,\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i))].$$

Finally, because $\mathcal{A} \subseteq \mathcal{B}$, we can conclude that $f(\mathcal{B} \cup \{\boldsymbol{x}\}) - f(\mathcal{B}) \neq 0 \Rightarrow f(\mathcal{A} \cup \{\boldsymbol{x}\}) - f(\mathcal{A}) \neq 0$ and therefore the aforementioned cases are sufficient. Combining all cases, we can see that the contribution of each individual to the marginal difference of f is always greater or equal under the set of counterfactual explanations $\mathcal{A}$ than under the set of counterfactual explanations $\mathcal{B}$. As a direct consequence, it follows that f is submodular. Additionally, we can easily see that this contribution is always greater or equal than zero, leading to the conclusion that f is also monotone.

B.3 Proof of Proposition 4

By definition, since $\mathcal{A} \subseteq \mathcal{P}_{\pi^*_{\mathcal{A}}}$, it readily follows that $\pi^*_{\mathcal{A}}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathcal{A}$. To find the remaining values of the decision policy, we first observe that, for each $\boldsymbol{x} \notin \mathcal{A}$, the value of the decision policy $\pi^*_{\mathcal{A}}(\boldsymbol{x})$ does not affect the best-responses of the individuals with initial feature values $\boldsymbol{x}' \neq \boldsymbol{x}$. As a result, we can just set $\pi^*_{\mathcal{A}}(\boldsymbol{x})$ for all $\boldsymbol{x} \notin \mathcal{A}$ independently for each feature value $\boldsymbol{x}$ such that the best-response of the respective individuals is the one that contributes maximally to the overall utility.

First, it is easy to see that, for all $\boldsymbol{x} \notin \mathcal{A}$ such that $P(y=1\,|\,\boldsymbol{x}) < \gamma$, we should set $\pi^*_{\mathcal{A}}(\boldsymbol{x}) = 0$. Next, consider the feature values $\boldsymbol{x} \notin \mathcal{A}$ such that $P(y=1\,|\,\boldsymbol{x}) \geq \gamma$. Here, we distinguish two cases. If there exists $\boldsymbol{x}' \in \mathcal{A}$ such that $c(\boldsymbol{x}, \boldsymbol{x}') \leq 1 \wedge P(y=1\,|\,\boldsymbol{x}') > P(y=1\,|\,\boldsymbol{x})$, then, if the individuals move to that $\boldsymbol{x}'$, the corresponding contribution to the utility will be higher. Moreover, the value of the decision policy that maximizes their region of adaption (and thus increases their chances of moving to $\boldsymbol{x}'$) is clearly $\pi^*_{\mathcal{A}}(\boldsymbol{x}) = 0$. If there does not exist $\boldsymbol{x}' \in \mathcal{A}$ such that $c(\boldsymbol{x}, \boldsymbol{x}') \leq 1 \wedge P(y=1\,|\,\boldsymbol{x}') > P(y=1\,|\,\boldsymbol{x})$, then, the contribution of the corresponding individuals to the utility will be higher if they keep their initial feature values. Moreover, the value of the decision policy that will maximize this contribution will be clearly $\pi^*_{\mathcal{A}}(\boldsymbol{x}) = 1$.

B.4 Proof of Proposition 5

It readily follows that the function h is non-negative from the fact that, if the decision maker is rational, $\pi(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$ such that $P(y=1\,|\,\boldsymbol{x}) < \gamma$.

Next, consider two sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{Y}$ such that $\mathcal{A} \subseteq \mathcal{B}$ and a feature value $\boldsymbol{x} \in \mathcal{Y} \setminus \mathcal{B}$. Also, let $\mathcal{E}_{\mathcal{S}}(\boldsymbol{x}_i)$ be the counterfactual explanation given to the individuals with initial feature value $\boldsymbol{x}_i$ under a set of counterfactual explanations $\mathcal{S}$. Then, it is clear that the marginal difference $h(\mathcal{S} \cup \{\boldsymbol{x}\}) - h(\mathcal{S})$ only depends on individuals with initial features $\boldsymbol{x}_i$ such that either $1 - c(\boldsymbol{x}_i, \boldsymbol{x}) \geq 0$ and $\boldsymbol{x} = \mathcal{E}_{\mathcal{S} \cup \{\boldsymbol{x}\}}(\boldsymbol{x}_i)$ or $\boldsymbol{x}_i = \boldsymbol{x}$. Moreover, if $1 - c(\boldsymbol{x}_i, \boldsymbol{x}) \geq 0$ and $\boldsymbol{x} = \mathcal{E}_{\mathcal{S} \cup \{\boldsymbol{x}\}}(\boldsymbol{x}_i)$, the contribution to the marginal difference is positive and, if $\boldsymbol{x}_i = \boldsymbol{x}$, the contribution to the marginal difference is negative.

Consider first the individuals with initial features $\boldsymbol{x}_i$ such that $1 - c(\boldsymbol{x}_i, \boldsymbol{x}) \geq 0$ and $\boldsymbol{x} = \mathcal{E}_{\mathcal{A}\cup\{\boldsymbol{x}\}}(\boldsymbol{x}_i)$. We can divide all of these individuals into three cases:

1. $\pi_{\mathcal{B}}(\boldsymbol{x}_i) = 0$: in this case, $\boldsymbol{x}_i \notin \mathcal{B}$ and the individuals change their best-response from $\mathcal{E}_{\mathcal{B}}(\boldsymbol{x}_i)$ to $\boldsymbol{x}$. Moreover, under the set of counterfactual explanations $\mathcal{A}$, their best-response is either $\boldsymbol{x}_i$ or $\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i)$ and it changes to $\boldsymbol{x}$. Then, using a similar argument as in the proof of proposition 2, we can conclude that the contribution of the individuals to the marginal difference is greater or equal under the set of counterfactual explanations $\mathcal{A}$ than under $\mathcal{B}$.
2. $\pi_{\mathcal{B}}(\boldsymbol{x}_i) = 1 \wedge \pi_{\mathcal{A}}(\boldsymbol{x}_i) = 0$: in this case, $\boldsymbol{x}_i \notin \mathcal{A}$ and $\boldsymbol{x}_i \in \mathcal{B}$. Therefore, under the set of counterfactual explanations $\mathcal{A}$, the individuals' best-response changes from $\mathcal{E}_{\mathcal{A}}(\boldsymbol{x}_i)$ to $\boldsymbol{x}$ and there is a positive contribution to the marginal difference while, under $\mathcal{B}$, the individuals' best response does not change and the contribution to the marginal difference is zero.
3. $\pi_{\mathcal{B}}(\boldsymbol{x}_i) = 1 \wedge \pi_{\mathcal{A}}(\boldsymbol{x}_i) = 1$: in this case, $\boldsymbol{x}_i \notin \mathcal{B}$. Therefore, the best-response changes from $\boldsymbol{x}_i$ to $\boldsymbol{x}$ under both sets of counterfactual explanations and there is an equal positive contribution to the marginal difference.

Now, consider the individuals with initial features $\boldsymbol{x}_i$ such that $\boldsymbol{x}_i = \boldsymbol{x}$. We can divide all of these individuals also into three cases:

1. $\pi_{\mathcal{A}}(\boldsymbol{x}) = \pi_{\mathcal{B}}(\boldsymbol{x}) = 0$: in this case, under both sets of counterfactual explanations, the counterfactual explanation $\boldsymbol{x}$ changes the value of the decision policy to $\pi_{\mathcal{A}\cup\{\boldsymbol{x}\}}(\boldsymbol{x}) = \pi_{\mathcal{B}\cup\{\boldsymbol{x}\}}(\boldsymbol{x}) = 1$. Moreover, the contribution to the marginal difference is less negative under the set of counterfactual explanations $\mathcal{A}$ than under $\mathcal{B}$ since $P(y = 1 \,|\, \mathcal{E}_{\mathcal{A}}(\boldsymbol{x})) \leq P(y = 1 \,|\, \mathcal{E}_{\mathcal{B}}(\boldsymbol{x}))$ and thus $P(\boldsymbol{x})[P(y = 1 \,|\, \boldsymbol{x}) - P(y = 1 \,|\, \mathcal{E}_{\mathcal{A}}(\boldsymbol{x}))] \geq P(\boldsymbol{x})[P(y = 1 \,|\, \boldsymbol{x}) - P(y = 1 \,|\, \mathcal{E}_{\mathcal{B}}(\boldsymbol{x}))]$.
2. $\pi_{\mathcal{A}}(\boldsymbol{x}) = 1 \wedge \pi_{\mathcal{B}}(\boldsymbol{x}) = 0$: in this case, under the set of counterfactual explanations $\mathcal{A}$, the individuals' best response does not change and thus the contribution to the marginal difference is zero and, under the set of counterfactual explanations $\mathcal{B}$, their best-response changes from $\mathcal{E}_{\mathcal{B}}(\boldsymbol{x})$ to $\boldsymbol{x}$ and thus there is a negative contribution to the marginal difference *i.e.*, $P(\boldsymbol{x})[P(y = 1 \,|\, \boldsymbol{x}) - P(y = 1 \,|\, \mathcal{E}_{\mathcal{B}}(\boldsymbol{x}))] < 0$.
3. $\pi_{\mathcal{A}}(\boldsymbol{x}) = \pi_{\mathcal{B}}(\boldsymbol{x}) = 1$: in this case, under both sets of counterfactual explanations, the individuals' best response does not change and thus the contribution to the marginal difference is zero.

As a direct consequence of the above observations, it readily follows that $h(\mathcal{A} \cup \{\boldsymbol{x}\}) - h(\mathcal{A}) \geq h(\mathcal{B} \cup \{\boldsymbol{x}\}) - h(\mathcal{B})$ and therefore the function h is submodular.

However, in contrast with Section 3, the function h is non-monotone since it can happen that the negative marginal contribution exceeds the positive one. For example, consider the following instance of the problem, where $\boldsymbol{x} \in \{1, 2, 3\}$ with $\gamma = 0.1$:

$$P(\boldsymbol{x}) = 0.1\,\mathbb{I}(\boldsymbol{x} = 1) + 0.8\,\mathbb{I}(\boldsymbol{x} = 2) + 0.1\,\mathbb{I}(\boldsymbol{x} = 3),$$
$$P(y = 1 \,|\, \boldsymbol{x}) = 1.0\,\mathbb{I}(\boldsymbol{x} = 1) + 0.5\,\mathbb{I}(\boldsymbol{x} = 2) + 0.4\,\mathbb{I}(\boldsymbol{x} = 3),$$

and

$$c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{bmatrix} 0.0 & 0.2 & 0.3 \\ 0.3 & 0.0 & 0.7 \\ 0.4 & 0.5 & 0.0 \end{bmatrix}.$$

Assume there is a set of counterfactual explanations $\mathcal{A} = \{1\}$. Then, the optimal policy is given by $\pi^*_{\mathcal{A}}(1) = 1, \pi^*_{\mathcal{A}}(2) = 0, \pi^*_{\mathcal{A}}(3) = 0$ inducing a movement from feature values $2, 3$ to feature value 1, giving a utility equal to 0.9. Now, add $\boldsymbol{x} = 2$ to the set of counterfactual explanations *i.e.*, $\mathcal{A} = \{1, 2\}$. Then, the optimal policy is given by $\pi^*_{\mathcal{A}}(1) = 1, \pi^*_{\mathcal{A}}(2) = 1, \pi^*_{\mathcal{A}}(3) = 0$ inducing a movement from feature value 3 to feature value 1, giving a lower utility, equal to 0.5. Therefore, the function h is non-monotone.

C Additional details on the standard greedy algorithm and the randomized algorithm by Buchbinder et al. (2014)

Algorithm 1 summarizes the standard greedy algorithm, which starts from a solution set $\mathcal{A} = \emptyset$ and it iteratively adds to $\mathcal{A}$ the counterfactual explanation $\boldsymbol{x} \in \mathcal{P}_\pi \setminus \mathcal{A}$ that provides the maximum marginal difference $f(\mathcal{A} \cup \{\boldsymbol{x}\}) - f(\mathcal{A})$

ALGORITHM 1: Standard greedy algorithm (Nemhauser et al., 1978)

Input: Ground set of counterfactual explanations $\mathcal{P}_\pi$, parameter k and utility function f
Output: Set of counterfactual explanations $\mathcal{A}$
$\mathcal{A} \leftarrow \varnothing$
while $|\mathcal{A}| \leq k$ **do**
    $\boldsymbol{x}^* \leftarrow \text{argmax}_{\boldsymbol{x} \in \mathcal{P}_\pi \setminus \mathcal{A}} f(\mathcal{A} \cup \{\boldsymbol{x}\}) - f(\mathcal{A})$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{\boldsymbol{x}^*\}$
end while
return $\mathcal{A}$

ALGORITHM 2: Randomized algorithm by Buchbinder et al. (2014)

Input: Ground set of counterfactual explanations $\mathcal{Y}$, parameter k and utility function f
Output: Set of counterfactual explanations $\mathcal{A}$
1: $\mathcal{A} \leftarrow \varnothing$
2: **while** $|\mathcal{A}| \leq k$ **do**
3:     $\mathcal{B} \leftarrow \text{GetTopK}(\mathcal{Y}, \mathcal{A}, f)$
4:     $\boldsymbol{x}^* \sim \mathcal{B}$
5:     $\mathcal{A} \leftarrow \mathcal{A} \cup \{\boldsymbol{x}^*\}$
6: **end while**
7: **return** $\mathcal{A}$

Algorithm 2 is just a randomized variation of the standard greedy algorithm. It starts from a solution set $\mathcal{A} = \emptyset$ and it iteratively adds one counterfactual explanation $\boldsymbol{x} \in \mathcal{Y} \setminus \mathcal{A}$. However, instead of greedily choosing the element $\boldsymbol{x}$ that provides the maximum marginal difference $h(\mathcal{A} \cup \{\boldsymbol{x}\}) - h(\mathcal{A})$, it sorts all the candidate elements with respect to their marginal difference (line 3) and picks one at random among the top k (line 4).

To enjoy a $1/e$ approximation guarantee, Algorithm 2 requires that there are $2k < m$ candidate feature values whose marginal contribution to any set is zero. In our problem, this can be trivially satisfied by adding $2k$ feature values $\boldsymbol{x}$ to $\mathcal{X}$ such that $P(y = 1 \,|\, \boldsymbol{x}) = \gamma$, $P(\boldsymbol{x}) = 0$ and $c(\boldsymbol{x}, \boldsymbol{x}_j) = c(\boldsymbol{x}_j, \boldsymbol{x}) = 2 \; \forall \boldsymbol{x}_j \in \mathcal{X}$. If the algorithm adds some of those counterfactual explanations to the set $\mathcal{A}$, it is easy to see that we can ignore them without causing any difference in utility or best-responses.

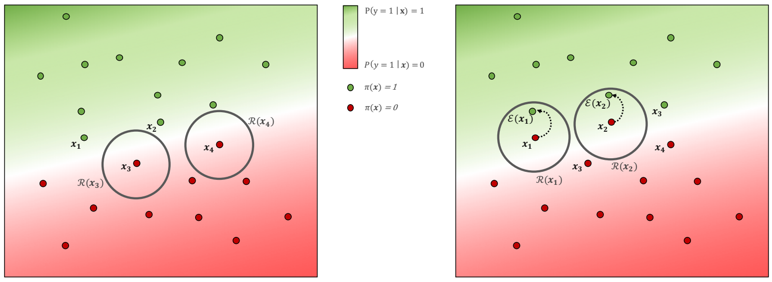
D Jointly optimizing the decision policy and the counterfactual explanations

Figure 5 shows that, by jointly optimizing both the decision policy and the counterfactual explanations, we may obtain an additional gain in terms of utility in comparison with just optimizing for the set of counterfactual explanations given the optimal decision policy in a non-strategic setting.

E Experiments on Synthetic Data

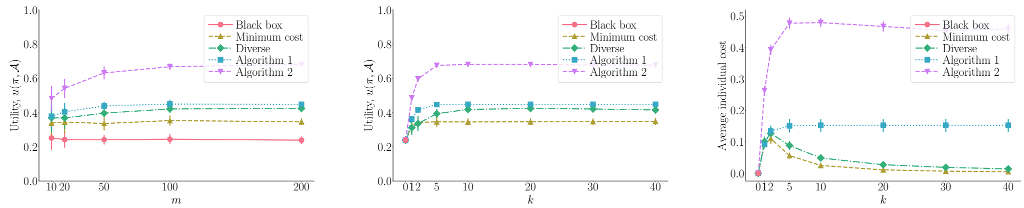
Experimental setup. For simplicity, we consider feature values $\boldsymbol{x} \in \{0, \ldots, m-1\}$ and $P(\boldsymbol{x} = i) = p_i / \sum_j p_j$ where p_i is sampled from a Gaussian distribution $N(\mu = 0.5, \sigma = 0.1)$ truncated from below at zero. We also sample $P(y = 1 \,|\, \boldsymbol{x}) \sim U[0, 1]$, $c(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim U[0, 1]$ for 50% of all pairs and $c(\boldsymbol{x}_i, \boldsymbol{x}_j) = 2$ for the rest. Finally, we set $\gamma = 0.3$. In this section, we compare the utility achieved by our explanation methods with the same baselines we used on real data.

Results. Figures 6(a,b) show the utility achieved by each of the decision policies and counterfactual explanations for several numbers of feature values m and counterfactual explanations k. We find several interesting insights: (i) the decision policies given by Eq. 5 and the counterfactual explanations found by Algorithm 2 beat all other alternatives by large margins across the whole spectrum, showing that jointly optimizing the decision policy and the counterfactual explanations offer clear additional gains; (ii) the counterfactual explanations found by Algorithms 1 and 2 provide higher utility gains as the number of feature values increases and thus the search space of counterfactual explanations becomes larger; and, (iii) a small number of counterfactual explanations is enough to provide significant gains in terms of utility with respect to the optimal decision policy without counterfactual explanations.



Non-strategic policy Strategic policy

Figure 5: Jointly optimizing the decision policy and the counterfactual explanations can offer additional gains. The left panel shows the optimal (deterministic) decision policy π under non-strategic behavior, as given by Eq. 6. Here, there does not exist a set of counterfactual explanations $\mathcal{A} \in \mathcal{P}_\pi$ that increases the utility of the policy. This happens because the area of adaption of $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ does not include any feature value that receives a positive decision. The right panel shows the decision policy and counterfactual explanations that are (jointly) optimal in terms of utility, as given by Eq. 4. Here, the individuals with feature values $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ receive $\mathcal{E}(\boldsymbol{x}_1)$ and $\mathcal{E}(\boldsymbol{x}_2)$, respectively, as counterfactual explanations. Since these explanations are within their areas of adaptation $\mathcal{R}(\boldsymbol{x}_1)$ and $\mathcal{R}(\boldsymbol{x}_2)$, they change their initial feature values in order to receive a positive decision.



(a) Utility vs. # feature values (b) Utility vs. # explanations (c) Individual cost vs. # explanations

Figure 6: Results on synthetic data. Panels (a) and (b) show the utility achieved by six types of decision policies and counterfactual explanations against the total number of feature values m and the number of counterfactual explanations k, respectively. Panel (c) shows the average cost individuals had to pay to change from their initial features to the feature value of the counterfactual explanation they receive under the same five types of decision policies and counterfactual explanations. In Panel (a), we set $k = 0.1m$ and, in Panels (b) and (c), we set $m = 200$. In all panels, we repeat each experiment 20 times.

Figure 6(c) shows the average cost individuals had to pay to change from their initial features to the feature value of the counterfactual explanation they receive. As one may have expected, the results show that, under the counterfactual explanations of minimum cost (Minimum cost and Diverse), the individuals invest less effort to change their initial features and the effort drops as the number of counterfactual explanations increases. In contrast, our methods incentivize the individuals to achieve the highest self-improvement, particularly when we jointly optimize the decision policy and the counterfactual explanations.

F Additional details on the experiments on real data

F.1 Feature representation & preprocessing steps

For each applicant in the lending dataset, the label y indicates whether an applicant fully pays a loan ($y = 1$) or ends up to a default/charge-off ($y = 0$) and the features $\boldsymbol{x}$ are:

- Loan Amount: The amount that the applicant initially requested.
- Employment Length: How long the applicant has been employed.

- Debt to Income Ratio: The ratio between the applicant's financial debts and her average income.
- FICO Score: The applicant's FICO score, which is a credit score based on consumer credit files. The FICO scores are in the range of 300-850 and the average of the high and low range for the FICO score of each applicant has been used for this study.
- Annual Income: The declared annual income of the applicant.

Here, we assume that all of the aforementioned features are *actionable*, meaning that an individual denied a loan can change their values in order to get a positive decision.

For each credit card holder in the credit dataset, the label indicates whether a credit card holder will default during the next month ($y = 0$) or not ($y = 1$) and the features $\boldsymbol{x}$ are:

- Marital status: Whether the person is married or single.
- Age Group: Group depending on the person's age (< 25, $25 - 39$, $40 - 59$, > 60).
- Education Level: The level of education the individual has acquired (1-4).
- Maximum Bill Amount Over Last 6 Months
- Maximum Payment Amount Over Last 6 Months
- Months With Zero Balance Over Last 6 Months
- Months With Low Spending Over Last 6 Months
- Months With High Spending Over Last 6 Months
- Most Recent Bill Amount
- Most Recent Payment Amount
- Total Overdue Counts
- Total Months Overdue

Here, we assume that all features except Marital Status, Age Group and Education Level are actionable and, among the actionable features, we assume that Total Overdue Counts and Total Months Overdue can only increase.

In both cases, note that the actionable features are numerical, however, our methodology only allows for discrete valued features. Therefore, rather than using the numerical values as features, we first cluster the loan applicants (or credit card holders) into k groups based on the original numerical features using k-clustering and then, for each applicant (or credit card holder), use the cluster identifier it belongs to, represented using a one-hot encoding, as a feature. After this preprocessing step, the discrete feature values $\boldsymbol{x}_i$ consists of all possible value combinations of discrete non-actionable features, if any, and cluster identifiers.

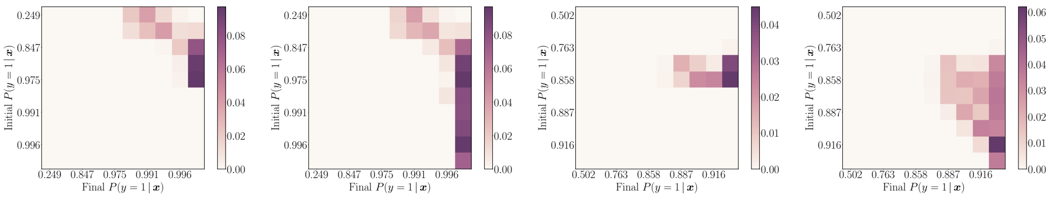
To approximate the values of the conditional distribution $P(y \,|\, \boldsymbol{x})$, we train four types of classifiers (Multi-layer perceptron, support vector machine, logistic regression, decision tree) using the default scikit-learn parameters and then choose the pair of classifier type and number of clusters k that maximizes accuracy, estimated using 5-fold cross validation. Finally, we set γ equal to the 50-th percentile of all the individuals' $P(y = 1 \,|\, \boldsymbol{x})$ values causing a 50% acceptance rate by the optimal threshold policy in the non strategic setting. Table F.1 summarizes the resulting experimental setup for both datasets.

F.2 Computation of the cost between feature values

To model each individual's best response, we need to estimate the cost between each pair of feature values $c(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Inspired by Ustun et al. (2019), we compute this cost based on the maximum percentile shift among actionable features. More specifically, let $\mathcal{L}$ be the set of actionable numerical features and $\bar{\mathcal{L}}$ be the set of non-actionable (discrete-valued) features and note that, in the credit dataset, $\bar{\mathcal{L}}$ contains Marital Status, Age Group and Education Level and $\bar{\mathcal{L}}$ contains the remaining features consist $\mathcal{L}$ and,

Table 1: Dataset details

Dataset	# of samples	Classifier	k	Accuracy	m	γ
credit	30000	Logistic Regression	100	80.4%	3200	0.85
lending	1266817	Logistic Regression	400	89.9%	400	0.97



(a) Alg. 1, lending (b) Alg. 2, lending (c) Alg. 1, credit (d) Alg. 2, credit

Figure 7: Transportation of mass induced by the policies and counterfactual explanations used in Algorithm 1 and 2 in both the lending and the credit dataset. For each individual in the population, whose best-response is to change her feature value, we record her outcome $P(y = 1 \mid \boldsymbol{x})$ before the best response (Initial $P(y = 1 \mid \boldsymbol{x})$) and after the best response (Final $P(y = 1 \mid \boldsymbol{x})$). In each panel, the color is proportional to the percentage of individuals who move from initial $P(y = 1 \mid \boldsymbol{x})$ to final $P(y = 1 \mid \boldsymbol{x})$ and we set $\alpha = 2$.

in the lending dataset, $\mathcal{L}$ contains all features. Then, for each pair of feature values $\boldsymbol{x}_i, \boldsymbol{x}_j$ we define the cost function as:

$$c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} \alpha \cdot \max_{l \in \mathcal{L}} |Q_l(x_{j,l}) - Q_l(x_{i,l})| & \text{if } x_{i,l} = x_{j,l} \ \forall l \in \bar{\mathcal{L}} \\ \infty & \text{otherwise,} \end{cases} \tag{7}$$

where $x_{j,l}$ is the value of the l-th feature for the feature value $\boldsymbol{x}_j$, $Q_l(\cdot)$ is the CDF of the numerical feature $l \in \mathcal{L}$ and $\alpha \geq 1$ is a scaling factor. As an exception, in the credit dataset, we always set the cost $c(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between two feature values to ∞ if $Q_l(x_{j,l}) < Q_l(x_{i,l})$ for $l \in \{\text{Total Overdue Counts}, \text{Total Months Overdue}\}$ considering the fact that history of overdue payments cannot be erased.

Finally, we would like to acknowledge that more sophisticated cost functions can be designed in terms of feasibility and difficulty of adaptation, taking into account domain knowledge and information about the deployed classifier, however, it goes beyond the scope of our work.

F.3 Transportation of mass

To measure the transportation of mass induced by the policies and counterfactual explanations used in Algorithm 1 and 2 in both the lending and the credit dataset, we proceed as follows. For each individual in the population whose best-response is to change her feature value, we record her outcome $P(y = 1 \mid \boldsymbol{x})$ before and after the best response. Then, we discretize the outcome values using percentiles. Figure 7 summarizes the results, which show several interesting insights. In the lending dataset, we observe that a large portion of individuals do improve their outcome even if we only optimize the counterfactual explanations (Panel (a)). In contrast, in the credit dataset, we observe that, if we only optimize the counterfactual explanations (Panel (c)), most individuals do not improve their outcome. That being said, if we jointly optimize the decision policy and counterfactual explanations (Panels (b) and (d)), we are able to incentivize a large portion of individuals to self improve in both datasets.

F.4 Examples of counterfactual explanations

In this section, we focus on the credit dataset and look more closely into the counterfactual explanations $\mathcal{E}_m(\boldsymbol{x})$ and $\mathcal{E}(\boldsymbol{x})$ provided by the minimum cost baseline and Algorithm 1, respectively, by means of an (anecdotal) example. To this end, for a fixed α and k, we first track down the individuals whose best-response under both methods is to change their initial features to the provided counterfactual

Table 2: Counterfactual explanations $\mathcal{E}_m(\boldsymbol{x})$ and $\mathcal{E}(\boldsymbol{x})$ provided by the minimum cost baseline and Algorithm 1, respectively, to an individual with initial feature value $\boldsymbol{x}$. Initially, the individual's outcome is $P(y = 1 \,|\, \boldsymbol{x}) = 0.84$ and, after best-response, her outcome is $P(y = 1 \,|\, \mathcal{E}_m(\boldsymbol{x})) = 0.87$ and $P(y = 1 \,|\, \mathcal{E}(\boldsymbol{x})) = 0.89$, respectively. In both methods, we set $\alpha = 2$ and $k = 160$.

Feature	$\boldsymbol{x}$	$\mathcal{E}_m(\boldsymbol{x})$	$\mathcal{E}(\boldsymbol{x})$
Married	No	No	No
Age group	Under 25	Under 25	Under 25
Education	Student	Student	Student
Maximum Bill Amount Over Last 6 Months	\$2246	\$2084	\$1929
Maximum Payment Amount Over Last 6 Months	\$191	\$188	\$221
Months With Zero Balance Over Last 6 Months	0	0	0
Months With Low Spending Over Last 6 Months	0	0	0
Months With High Spending Over Last 6 Months	4	2	1
Most Recent Bill Amount	\$2145	\$2003	\$1750
Most Recent Payment Amount	\$123	\$124	\$100
Total Overdue Counts	0	0	0
Total Months Overdue	0	0	0

explanation. Then, for each of these individuals, we compare the counterfactual explanations provided by each of both methods.

Table F.4 shows the initial features $\boldsymbol{x}$ together with the counterfactual explanations $\mathcal{E}_m(\boldsymbol{x})$ and $\mathcal{E}(\boldsymbol{x})$ for one of the above individuals picked at random. In this example, the individual is a university student, unmarried and under the age of 25 who is advised to follow the counterfactual explanations to maintain her credit. Since the marital status, age group and level of education are all non-actionable features, both counterfactual explanations maintain the initial values for those features. Under the minimum cost baseline, the bank would advise the individual to reduce her monthly credit card bill by ∼\$150 and limit high spending to 2 months per semester so that her risk of default would decrease from 16% to 13%. However, under Algorithm 1, the bank would advise to reduce her monthly credit card bill by ∼\$400, limit high spending to 1 month per semester, and additionally increase her monthly credit card payoff slightly so that her risk of default would decrease to 11%. Since by construction, both $\mathcal{E}_m(\boldsymbol{x})$ and $\mathcal{E}(\boldsymbol{x})$ are inside the region of adaptation of $\boldsymbol{x}$, the individual is guaranteed to follow the advise in both cases, however, under Algorithm 1, the individual would be less likely to default and achieve a superior long-term well being.