The Risks of Recourse in Binary Classification

Hidde Fokkema
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
h.j.fokkema@uva.nl

Damien Garreau
CAIDAS
University of Würzburg
damien.garreau@uni-wuerzburg.de

Tim van Erven
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
tim@timvanerven.nl

Abstract

Algorithmic recourse provides explanations that help users overturn an unfavorable decision by a machine learning system. But so far very little attention has been paid to whether providing recourse is beneficial or not. We introduce an abstract learning-theoretic framework that compares the risks (i.e., expected losses) for classification with and without algorithmic recourse. This allows us to answer the question of when providing recourse is beneficial or harmful at the population level. Surprisingly, we find that there are many plausible scenarios in which providing recourse turns out to be harmful, because it pushes users to regions of higher class uncertainty and therefore leads to more mistakes. We further study whether the party deploying the classifier has an incentive to strategize in anticipation of having to provide recourse, and we find that sometimes they do, to the detriment of their users. Providing algorithmic recourse may therefore also be harmful at the systemic level. We confirm our theoretical findings in experiments on simulated and real-world data. All in all, we conclude that the current concept of algorithmic recourse is not reliably beneficial, and therefore requires rethinking.

1 Introduction

Machine learning (ML) models are increasingly being used to make consequential decisions in areas such as finance (Mukerjee et al., 2002), healthcare (Begoli et al., 2019; Grote and Berens, 2020), and hiring (Nabi and Shpitser, 2018; Schumann et al., 2020). When these decisions are unfavorable to the people they affect, algorithmic recourse provides explanations and recommendations to favorably change their situation (Karimi et al., 2022). For instance, when an individual is denied a bank loan, they might like to know the reasons and in particular what they can do to get a loan in the future.

A prominent approach to providing recourse is via counterfactual explanations, which suggest how the individual should change their features in order to flip the decision of the ML model (Wachter et al., 2017; Ustun et al., 2019; Joshi et al., 2019). Originally, counterfactuals were chosen to minimize the distance between the original and the new features (Wachter et al., 2017), but more recently attention has also been paid to generating realistic suggestions which are actionable and lie on the data manifold (Ustun et al., 2019; Joshi et al., 2019). In addition, various types of robustness have been studied, including to random perturbations (Virgolin and Fracaros, 2023; Dominguez-Olmedo et al., 2022; Pawelczyk et al., 2022b), to data shifts (Rawal et al., 2020; Dutta et al., 2022), or to the case that the counterfactual might not be perfectly implementable (Artelt et al., 2021). It has further been recognized that providing recourse has consequences at the population level, because it changes the distribution of the data. These consequences have been studied in the context of fairness for subgroups (Gupta et al., 2019) and with respect to social segregation (Gao and Himabindu, 2023), but so far there has been no work that studies the consequences of providing recourse for classification accuracy.

To see why accuracy matters, consider again the loan example mentioned above. If a person is able to repay a loan they got through recourse, then recourse has been beneficial. But if they end up defaulting on their payment, then recourse has actually been harmful, both for the user and the lending institution. Providing recourse in a way that undermines the accuracy of the ML model in determining which users are likely to default, can therefore be dangerous. In fact, the bank loan example above, which is standard in the recourse literature, is also used as a motivating example in the context of strategic classification. There, it is seen as a significant risk that loan applicants might try to game the system by changing their features to flip the class without actually improving their true creditworthiness (Brown et al., 2022; Perdomo et al., 2020; Milli et al., 2019).

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

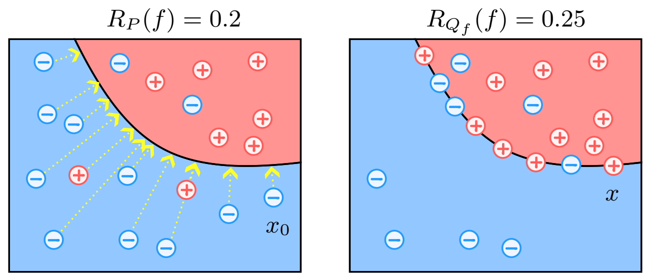

Figure 1: *Left panel:* Initial situation, the ML model classifies individual with starting features x_0 either negatively (in blue) or positively (in red). Its risk is denoted by $R_P(f)$. Points classified negatively are given the opportunity to move to the decision boundary (yellow dotted arrows). *Right panel:* The points close enough to the boundary accept recourse and move towards the decision boundary. The risk with recourse, $R_{Q_f}(f)$, is then higher, because at the decision boundary the uncertainty about the true class is maximal, and the points that accepted recourse are now more likely to be misclassified.

Main Contributions In this work, we study the effects of recourse on the classification accuracy at the population level. All our results are obtained in the context of a new learning-theoretic framework, which we introduce in Section 2. Accuracy is measured by the risk, which is the expected loss of a given classifier. When recourse is provided, it changes the distribution of the data, and hence the risk. We are primarily interested in whether recourse makes the risk go up or down. To answer this question it matters how the class probabilities of the users change upon receiving recourse. We distinguish between the *compliant* case, in which these class probabilities truly improve, and the *defiant* case, in which the class probabilities do not improve at all (for instance because the users are trying to game the system). In Section 3 we show that, if the classifier is optimal without recourse, then recourse will be harmful, because it increases the risk both for the compliant and for the defiant case. The reason is that recourse pushes users towards the decision boundary, where the class uncertainty is higher, which therefore leads to more mistakes. See Figure 1 for an illustration. Section 4 extends these results to probabilistic classifiers that are only near-optimal, which allows for estimation errors, and to surrogate losses like, e.g., the cross-entropy loss. In Section 5 we recognize that the party deploying the classifier may strategically choose their classifier in order to minimize the resulting risk after providing recourse. We obtain separate results for the defiant and compliant cases, which show that there is an incentive to preemptively undo the effect of recourse. For the defiant case, this makes the risks with and without recourse identical, so implementing recourse only places a burden on all parties, without any resulting advantage. For the compliant case, the risk with recourse does decrease, so this is the only case where we do observe an advantage to providing recourse. Finally, in Section 6 we corroborate our theoretical results by experiments, in which we observe the risk increase for a large majority of the experiments both on synthetic data and on real data. We also provide the code that produced the results of our experiments as a GitHub repository.[1]

Not Reliably Beneficial In summary, our findings show that there are many common cases in which recourse is harmful, because it leads to worse classification accuracy. This suggests that instead of debating how to provide recourse, we should rethink whether the current approach to recourse is desirable at all. Notably, there is no escape by pointing to exceptions in which recourse is beneficial, e.g., our results on strategic classification for the compliant case, or by pointing to specific examples where it is beneficial in practice: if recourse is not reliably beneficial nearly all the time, then it is not suitable to be broadly adopted.

1.1 Further Related Work

Causality and Algorithmic Recourse The difference between the defiant and the compliant case has already been noted in the causal algorithmic recourse community. This has lead to counterfactual methods with guarantees for actual improvement of the class probabilities, which ensure that we are closer to the compliant case König et al. (2021); König et al. (2023). However, our results show that, even in the fully compliant case, recourse may still be harmful. Such unexpected harmful effects of well-intended interventions have also been found in the context of fairness (Liu et al., 2018). More generally, it has been pointed out that users can only act on counterfactual recommendations if these take the causal relation between the user's actions and their features into account (Karimi et al., 2021). Our framework is general enough to express such causal interventions, because they only affect the risk via their effect on the distribution of the data.

Strategic Classification Strategic classification considers the effect of deploying a classifier in an environment with strategic players, who want to change their features in order to influence how they are classified (Hardt et al., 2016; Levanon and Rosenfeld, 2021; Miller et al., 2020; Tsirtsis et al., 2019; Yatong Chen, 2021). This makes the distribution of the data dependent on f as well, because the behavior of the players depends on the classifier f. The more abstract setting in which there can be any dependence between f and the data distribution, has been studied under the heading of performative prediction (Perdomo et al., 2020; Mofakhami et al., 2023). Our results about strategizing in Section 5 are a special case of strategic classification, in which the behavior of the players is guided by the recourse mechanism. In contrast to previous results that mostly considered how to minimize the risk in f while taking the dependence of f

[1] github.com/HiddeFok/consequences-of-recourse

on the distribution into account, our aim is to quantify the difference in the risk when we compare the settings with and without recourse.

2 Framework and Main Definitions

In this section we formalize the effect of recourse by comparing the risk in the situation without recourse to the risk with recourse applied.

2.1 General Framework

We consider binary classification, in which users with corresponding features x from a closed, convex domain $\mathcal{X} \subseteq \mathbb{R}^d$ will be classified into classes $\mathcal{Y} = \{-1, +1\}$. We assume a model $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ has already been trained. This may be a deterministic classifier, with $\widehat{\mathcal{Y}} = \{-1, +1\}$, or a probabilistic classifier, with $\widehat{\mathcal{Y}} = [0, 1]$, for which $f(x)$ represents the probability that x should be classified as $+1$. The error of a prediction $\hat{y} \in \widehat{\mathcal{Y}}$ with respect to the true label $y \in \mathcal{Y}$ is measured by a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$. For instance, for deterministic predictions $\hat{y} \in \{-1, +1\}$, the $0/1$ loss is $\ell(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$, and, for probabilistic predictions $\hat{y} \in [0, 1]$, the log loss or cross-entropy loss is $\ell(\hat{y}, y) = \frac{1}{2}(1 + y) \ln \frac{1}{\hat{y}} + \frac{1}{2}(1 - y) \ln \frac{1}{1-\hat{y}}$.

In the absence of recourse, the data will consist of pairs (X_0, Y) from $\mathcal{X} \times \mathcal{Y}$ with distribution P, and the quality of f is evaluated by its risk

$$R_P(f) = \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0), Y)]. \quad \text{(Risk without Recourse)}$$

A classifier $f_P^* \in \arg\min_f R_P(f)$, which minimizes the risk, is called Bayes-optimal. For instance, for $0/1$ loss, $f_P^*(x_0) = \text{sign}(P(Y = 1|X_0 = x_0) - \frac{1}{2})$ is Bayes-optimal. Throughout the paper, we take the sign function $\text{sign}(z)$ to be $+1$ if $z \geq 0$ and -1 for $z < 0$.

When we add recourse to the mix, a user first arrives with feature vector X_0, which is drawn according to the marginal distribution of P on $\mathcal{X}$. Then, depending on the original features X_0, the specifics of the recourse protocol, and the model f, the user's features are transformed into new features $X \in \mathcal{X}$. Here, X may be a deterministic function of X_0, but in general it can also depend on X_0 in a non-deterministic way if the recourse protocol is randomized or when the user's response to recourse is not fully predictable. Finally, a label Y is generated, and we let Q_f denote the resulting distribution of (X_0, X, Y). The resulting risk is then measured under the marginal distribution of (X, Y) under Q_f:

$$R_{Q_f}(f) = \mathop{\mathbb{E}}_{(X,Y)\sim Q_f}[\ell(f(X), Y)]. \quad \text{(Risk with Recourse)}$$

Thus, the marginal distribution of X_0 under Q_f is always the same as under P. Note further that f influences the risk with recourse in two ways: directly via its predictions $f(X)$ and indirectly via its effect on the distribution Q_f. Except for Section 5 we will think of f as fixed, and we will simplify notation by writing Q instead of Q_f.

As motivated in the introduction, we care about the accuracy of classifiers at the population level. This is measured by the risk, so we will say that recourse is beneficial if the risk under Q is smaller than the risk under P, and harmful otherwise.

2.2 Specializing the Framework

The framework above is so general that it can represent any mechanism for providing recourse. In order to say something concrete, we have to specialize it further.

Effect on the Label Distribution Naively, we might expect that changing the user's features from X_0 to X would also change their label distribution from $P(Y|X_0)$ to $P(Y|X)$, but what actually happens depends on the underlying causal effect of providing recourse (Miller et al., 2020; König et al., 2023), and in general any effect on the label distribution is possible. We will focus on two extreme cases which differ in whether individuals fully comply with or fully defy this naive expectation:

(Compliant) $Q(Y \mid X_0, X) = P(Y \mid X)$. The change in features causes a true change in label probability.

(Defiant) $Q(Y \mid X_0, X) = P(Y \mid X_0)$. The user only changes their features, without altering their label probability.

We state all our results in terms of these two extreme cases. However, those results can easily be generalized to an intermediate setting by taking a convex combination of the compliant and defiant measure. The convex combination would then carry over towards the theoretical results.

The defiant case has also been referred to as "gaming"[2] (König et al., 2023; Perdomo et al., 2020). It is illustrated well by the following example by König et al. (2023): consider a classifier which classifies whether a patient is infected with Covid based on their symptoms. Then, taking cough drops to suppress coughing may change the classification without changing the true probability of being infected. This behaviour could also appear when there is no recourse considered. In that case, we assume that it is already modelled in the distribution P. In our setting, the act of giving recourse is what gives the users the opportunity to "game" the system, when they were not before. It should not come as a complete surprise that our results show that the risk increases when giving recourse in the defiant case. The more surprising conclusions are that even in the compliant case it is possible to observe a risk increase and that strategizing

[2] We avoid this terminology in the context of algorithmic recourse, because users may follow a recourse recommendation in good faith and still not change their label probability.

against the risk increase in the defiant case comes with its own negative consequences.

Recourse Mechanism We will think of class $+1$ as being favorable to the users, while class -1 is undesirable to them. For instance, $+1$ might represent a bank loan being granted, while -1 means that the loan application is rejected. Whenever a user with features X_0 is classified as $f(X_0) = -1$ by a deterministic classifier, they may request recourse. Many prominent approaches (Wachter et al., 2017; Ustun et al., 2019; Karimi et al., 2020; Pawelczyk et al., 2022a) to algorithmic recourse provide the user with a counterfactual explanation $X_0^{\text{CF}} = \varphi(X_0)$ which is the solution to an optimization problem of the form

$$X_0^{\text{CF}} \in \underset{z \in \mathcal{X}:\ f(z)=+1}{\arg\min}\ c(X_0, z), \tag{1}$$

where $c(x_0, z)$ models the cost for the user of moving from x_0 to z. This can describe many different cost mechanisms, and can even be used to express constraints like monotonicity in an Age feature or consistency with a causal model, by assigning large cost to any z that violates the constraints. For the optimization problem in (1) to be well-defined, we need to assume that the set

$$\{x \in \mathcal{X} \mid f(x) = +1\} \qquad \text{is closed.} \tag{2}$$

A consequence of this, is that a point on the decision boundary of a classifier will be classified as class $+1$. So, in order for f_P^* to satisfy this condition for $0/1$ loss, it matters that we defined $\text{sign}(0) = +1$ above. For many of our results, we will further assume that for any points x_0 and x the cost

$$z \mapsto c(x_0, z) \quad \text{increases monotonically on the line segment from } x_0 \text{ to } x, \tag{3}$$

which means that larger changes require more effort from the user. Under this assumption, φ always maps users x_0 in the negative class to the decision boundary; for users in the positive class, recourse does not do anything and $\varphi(x_0) = x_0$. (See Lemma 6 in Appendix A.) If the user implements the counterfactual explanation exactly, then $X = X_0^{\text{CF}}$, but they might also deviate from it in a stochastic way, which would make X a noisy approximation of X_0^{CF} (Pawelczyk et al., 2022b). For simplicity, we will focus on the noiseless case with $X = X_0^{\text{CF}}$. We do explicitly take into account the fact that not all users might receive recourse and that each user has a choice in whether to implement it. Let $B \in \{0, 1\}$ be an indicator variable for whether recourse is received and implemented, with conditional probability $\Pr(B = 1 \mid X_0) = r(X_0)$. It then follows that

$$X = (1 - B)X_0 + BX_0^{\text{CF}} = (1 - B)X_0 + B\varphi(X_0)\,.$$

Note that, when $f(X_0) = +1$, we always have $X = X_0$ irrespective of B, because $\varphi(X_0) = X_0$ as mentioned above. Some examples of possible r functions are:

- $r(x_0) = 1$. All users implement recourse;
- $r(x_0) = \mathbb{1}\{\|x_0 - \varphi(x_0)\| \leq D\}$ for some $D > 0$. Only those users within distance D of the decision boundary implement recourse;
- $r(x_0) = e^{-\frac{\|x_0 - \varphi(x_0)\|^2}{2\sigma^2}}$ for some $\sigma^2 > 0$. All users implement recourse with some probability and that probability is exponentially decreasing in the squared distance they have to cover, with a bandwidth σ^2.

More Complex Counterfactuals In our setting, we define a counterfactual explanation to be a solution to an optimization problem (1). In accordance with the early counterfactual methods developed in (Wachter et al., 2017; Karimi et al., 2020; Laugel et al., 2018). For the Compliant case, some of the recent counterfactual methods (Laugel et al., 2019; Kanamori et al., 2020; Parmentier and Vidal, 2021) can have more complex cost functions, with the goal of generating more realistic, feasible and robust counterfactuals. These methods produce counterfactuals that do not lie on the decision boundary, but may lie further into the positive class. For these methods, the cost function still has a component that depends on the Euclidean distance. The other component of the cost function depends on something called the outlier score, which indeed does not satisfy our requirement in Equation (3). However, depending on how these terms are balanced, the counterfactual point will still be close to the decision boundary, so the setting can still be approximated by the cases that we study.

Moreover, these methods are not designed to pay attention to the class probability $P(Y = +1|X)$, which is crucial to circumvent our result, and may therefore still produce counterfactuals for which $P(Y = +1|X) \approx \frac{1}{2}$. For instance, the methods in (Kanamori et al., 2020; Parmentier and Vidal, 2021) do pay attention to the marginal distribution of X, but not to $P(Y = +1|X)$. Consequently, all current robustness metrics or outlier metrics that do not increase the risk after providing recourse will do so by luck, but not by design.

We also like to remark that the results of the Defiant case will hold, whatever the cost function may be. This is because the conditional probability of being in the positive class does not change in that setting.

3 Risk Increase for the Bayes-Optimal Classifier

In this section we present our first main result, which relates the risk with recourse under Q to the risk without recourse under P. The result implies that the risk with recourse is larger, because recourse will move data from a region where the prediction is relatively certain, for example $P(Y = -1|X_0) = 0.9$, to the decision boundary, where things are the least certain, because $P(Y = +1|X) = 1/2$. We also illustrate this in an example with Gaussian data. The proofs

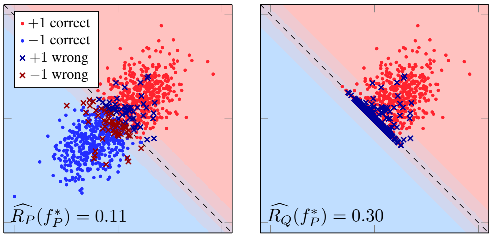

Figure 2: *Left:* Bayes classifier, original predictions; *Right:* predictions after providing recourse in the compliant case.

and additional details for the example in this section can be found in Appendix B.

Theorem 1 (Bayes-Optimal Classifier Risk Increase). *Let ℓ be the $0/1$ loss, and assume the setting of Section 2.2 (i.e., (1), (2), (3)). Suppose that $P(Y=1|X_0=x)=\frac{1}{2}$ for all x on the decision boundary of f_P^*. Then*

(a) For the defiant case,

$$\begin{aligned} R_Q(f_P^*) &= P(B=1, f_P^*(X_0)=-1, Y=-1) \\ &\quad - P(B=1, f_P^*(X_0)=-1, Y=+1) \\ &\quad + R_P(f_P^*) \\ &\geq R_P(f_P^*); \end{aligned} \tag{4}$$

(b) For the compliant case,

$$\begin{aligned} R_Q(f_P^*) &= \tfrac{1}{2}P(B=1, f_P^*(X_0)=-1) \\ &\quad - P(B=1, f_P^*(X_0)=-1, Y=1) \\ &\quad + R_P(f_P^*) \\ &\geq R_P(f_P^*). \end{aligned} \tag{5}$$

Both inequalities are strict if $P(B=1, f_P^*(X_0)=-1) > 0$, i.e., if the probability of recourse in the negative class is non-zero.

Theorem 1 gives an explicit expression for the risk with recourse when f_P^* is the Bayes classifier for P. Under very general conditions, it shows that providing recourse always increases the risk, for any recourse probability function r and any monotonically increasing cost function c!

3.1 Gaussian Example

We proceed with a simple example that can be analyzed in closed form and plotted visually. We assume the data is generated as follows. Let $P(X_0 \mid Y=y)$ be $\mathcal{N}(\mu,\Sigma)$ for $y=+1$ and $\mathcal{N}(\nu,\Sigma)$ for $y=-1$ for positive definite Σ, with equal prior class probabilities $P(Y=-1) = P(Y=+1) = \frac{1}{2}$. For simplicity, we will assume that $\|\mu\|_{\Sigma^{-1}} = \|\nu\|_{\Sigma^{-1}}$, where $\|\mu\|^2_{\Sigma^{-1}} = \langle \mu, \Sigma^{-1}\mu\rangle$ and set $\theta := \Sigma^{-1}(\mu-\nu)$. Then, the optimal classifier is known to be $f_P^*(x_0) = \operatorname{sign}(x_0^\top\theta)$, and the Bayes risk can be expressed in terms of the distribution function Φ of a standard normal distribution: $R_P(f_P^*) = \Phi(-\frac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}})$. For Euclidean cost $c(x_0,z) = \|x_0 - z\|$, providing recourse boils down to projecting onto the hyperplane $\{x\in\mathcal{X} \mid x^\top\theta = 0\}$ and this projection can be expressed analytically by a linear transformation $\varphi(x_0) = \left(I - \frac{\theta\theta^\top}{\|\theta\|^2}\right)x_0$.

We see the effect of providing recourse on the data distribution and the risk for the compliant case in Figure 2. We have taken $\mu = (+1,+1)^\top$, $\nu = (-1,-1)^\top$ and $\Sigma = \left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right)$, and set $r(x_0)=1$. In this case, $R_P(f_P^*) = \Phi(-\frac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}}) \approx 0.1$. The figure also shows empirically that the risk increases, which matches the prediction by Theorem 1 that $R_Q(f_P^*) = \frac{1}{4} + \frac{1}{2}\Phi(-\frac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}}) \approx 0.31$. The defiant case is not shown, because it would result in a similar picture, but with $R_Q(f_P^*) = \frac{1}{2}$.

4 Risk Increase for Probabilistic Classifiers

In practice, we do not have direct access to the Bayes-optimal classifier and the classifier is learned from data. In this section, we therefore drop the requirement that the classifier is exactly Bayes-optimal. We will further consider probabilistic classifiers $g : \mathcal{X} \to [0,1]$. Thresholding g then leads to a binary classifier $f(x) = \operatorname{sign}(g(x) - \frac{1}{2})$. We will compare the risk with recourse to the risk without recourse, first for the $0/1$ loss and then for a class of surrogate losses that includes the cross-entropy loss. The assumptions we make differ, but in both cases the conclusion is that the risk with recourse exceeds the risk without recourse when g is sufficiently accurate. The proofs for this section are presented in Appendix C.

4.1 Risk Increase for the $0/1$ loss

We again focus on the $0/1$ loss first. We can handle the defiant case without further assumptions. But for the compliant case we require that g is highly accurate in the sense that its decision boundary is close to Bayes-optimal. A simple sufficient requirement would be that there exists $\varepsilon \geq 0$ such that

$$\left|\tfrac{1}{2} - P(Y=1 \mid X_0 = x)\right| \leq \varepsilon \quad \text{for all } x \text{ such that } g(x) = 1/2. \tag{A}$$

This gives a uniform control over deviations anywhere along the decision boundary of g. At the cost of a slightly more complicated condition, this uniform bound can be relaxed to an average under the distribution over the points from the negative class that get mapped to the decision boundary of g:

$$\int_{\{x_0 : g(x_0) < 1/2\}} |\tfrac{1}{2} - P(Y=1 \mid X = \varphi(x_0))| P(\mathrm{d}x_0) \leq \varepsilon. \tag{B}$$

Assuming g is continuous, it will equal $g(x) = 1/2$ for all points x on its decision boundary. When φ maps all points x_0 from the negative class to the decision boundary of g, it follows that (A) implies (B).

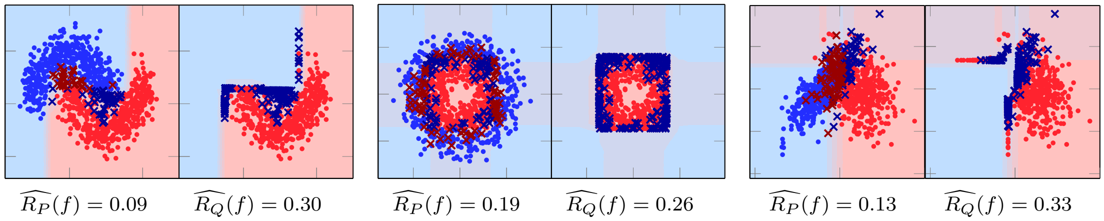

Figure 3: From left to right: Moons, Circles and Gaussian datasets. The left image for each shows the classifications with gradient boosted trees; the right image shows the effect of giving recourse.

We are now ready to derive an analogous result to Theorem 1:

Theorem 2 (Probabilistic Classifier Risk Increase, $0/1$ loss). *Let ℓ be the $0/1$ loss. Let $g : \mathcal{X} \to [0,1]$ be a continuous, probabilistic classifier, and define $f(x) = \operatorname{sign}(g(x) - \frac{1}{2})$. Assume (1), (2), (3) from Section 2.2. Then,*

(a) For the defiant case,

$$R_Q(f) = P(B = 1, f(X_0) = -1, Y = -1) - P(B = 1, f(X_0) = -1, Y = +1) + R_P(f). \quad (6)$$

Moreover, $R_Q(f) \geq R_P(f)$ if and only if

$$P(Y = -1 \mid B = 1, f(X_0) = -1) \geq \tfrac{1}{2}. \quad (7)$$

If we additionally assume that g satisfies (B) with $0 \leq \varepsilon \leq \frac{1}{2}$, then

(b) For the compliant case, $R_Q(f)$ is lower and upper bounded by

$$(\tfrac{1}{2} \pm \varepsilon)P(B = 1, f(X_0) = -1) + P(f(X_0) = +1, Y = -1) + P(B = 0, f(X_0) = -1, Y = 1). \quad (8)$$

Moreover, $R_Q(f) \geq R_P(f)$ if

$$P(Y = -1 \mid B = 1, f(X_0) = -1) \geq \tfrac{1}{2} + \varepsilon. \quad (9)$$

Equations (7) and (9) express that the class -1 is actually more likely (with a margin of ε) conditional on the set of points in the negative class that accept recourse. This will be satisfied when f is a reasonably accurate classifier. The intuition is that in this case moving points to the decision boundary is harmful, because they are more likely to be misclassified there. We also note that, for $\varepsilon = 0$, f will be equal to the Bayes-optimal classifier, and the condition is always satisfied, so we recover the conclusion from Theorem 1 that the risk will always increase.

4.2 Risk Increase for Surrogate Losses

In this section, we investigate the scenario in which the loss is not the $0/1$ loss, but rather a surrogate loss. We are primarily thinking of the cross-entropy loss, as defined in Section 2, but our result also applies to any other loss for probabilistic predictions $\hat{y} \in [0,1]$ which is such that $\ell(1/2, -1) = \ell(1/2, +1)$ is constant.

Theorem 3 (Probabilistic Classifier Risk Increase, Surrogate Loss). *Let $\ell : [0,1] \times \{-1,+1\} \to \mathbb{R}$ be any loss such that $\ell(1/2, -1) = \ell(1/2, +1) = c$ for some constant c. Let $g : \mathcal{X} \to [0,1]$ be a continuous, probabilistic classifier, and define $f(x) = \operatorname{sign}(g(x) - \frac{1}{2})$. Further assume (1), (2), (3) from Section 2.2. Then, both for the defiant and for the compliant case, we have $R_Q(g) \geq R_P(g)$ if and only if*

$$\mathop{\mathbb{E}}_{P}\left[\ell(g(X_0), Y) \mid f(X_0) = -1, B = 1\right] \leq c. \quad (10)$$

Condition 10 means that, on average over users from the negative class who receive recourse, the loss should be lower than the value of the loss at the decision boundary. This means that g should be a reasonably accurate classifier, which performs better on this group than simply predicting $1/2$. But it is much weaker than requiring that g should be close to Bayes-optimal, as we did in Theorems 1 and 2. We can get away with this weaker requirement, because, at the decision boundary, $g(x) = 1/2$ and therefore the loss is c regardless of the underlying distribution of Y. This is also the reason that the defiant and the compliant case coincide.

5 Strategic Classification

So far we have assumed that the classifier f was fixed, but when the party deploying f knows in advance that they will need to provide recourse, they have an incentive to strategically choose f in order to minimize the resulting risk under Q. In this section, we study the result of strategizing for both the defiant and compliant scenario. Before presenting our results, we first introduce the part of the setup that is common to both. At the end of the section, we reflect on our findings in a short discussion.

5.1 Common Setup

Throughout this section we focus on binary classifiers $f : \mathcal{X} \to \{-1,+1\}$ with the $0/1$ loss. And, since f is now variable, we write Q_f, φ_f and r_f instead of Q, φ and r. We assume that anyone either accepts or rejects recourse

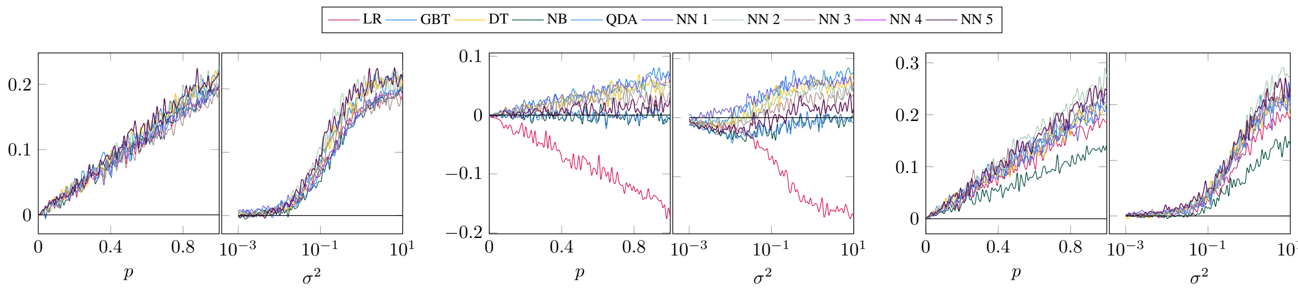

Figure 4: From left to right: Moons, Circles and Gaussian datasets. The left image for each shows the risk difference when $p \in [0,1]$; the right image shows the risk difference when $\sigma^2 \in [10^{-3}, 10^1]$ on a logarithmic scale

deterministically, i.e., that $r_f(x_0) \in \{0,1\}$ for all x_0. And we also assume that the classifier f is selected from a restricted class of functions $\mathcal{F}$. For notational sake, define $\varphi_f^r(x_0) = r_f(x_0)\varphi_f(x_0) + (1 - r_f(x_0))x_0$. Under the effect of recourse, $\mathcal{F}$ transforms into

$$\mathcal{F}_\varphi^r := \{x_0 \mapsto f(\varphi_f^r(x_0)) \mid f \in \mathcal{F}\}\,.$$

We say that $\mathcal{F}$ is *invariant under recourse* if, for any $f \in \mathcal{F}$, there exists a unique $f' \in \mathcal{F}$ such that f' with recourse is equivalent to f without recourse, i.e., $f'(\varphi_f^r(x_0)) = f(x_0)$ for all x_0. This implies, in particular, that $\mathcal{F}_\varphi^r = \mathcal{F}$. As a concrete example, one can think of linear classification, with recourse defined as bringing any point within distance less than $D > 0$ of the decision boundary to the positive class. In this example, shifting the original classifier by D orthogonally to the decision boundary in the direction of the positive class gives another equivalent classifier: it is thus invariant under recourse. Details for this example and another one are provided in Appendix D.

5.2 Defiant Case

In the defiant case, the setting above implies that providing recourse does not change the risk:

Theorem 4 (Strategizing in the Defiant Case). *Let ℓ be the $0/1$ loss, assume (1), (2), (3) from Section 2.2 with $r(x_0) \in \{0,1\}$ for all $x_0 \in \mathcal{X}$, and suppose $\mathcal{F}$ is invariant under recourse. Then, providing recourse in the defiant case does not change the risk when the party deploying the classifier strategizes to minimize their risk over $\mathcal{F}$:*

$$\min_{f\in\mathcal{F}} R_{Q_f}(f) = \min_{f\in\mathcal{F}} R_P(f)\,.$$

Intuitively, the reason is that in the defiant case it is strategically optimal to maintain the original decision boundary, because users do not really change upon receiving recourse. This is possible when $\mathcal{F}$ is recourse invariant, because then there is always a function available that compensates for the effect of recourse. Recourse therefore has no effect on the final decisions, but instead only places a burden on users who have to implement it and on the party deploying the classifier, which has to provide a recourse mechanism. In this case, recourse therefore has only negative effects, and may be considered harmful. We prove Theorem 4 in Appendix D.

5.3 Compliant Case

In the compliant case, the situation is different and strategizing can actually improve the risk. We require the following definition.

Definition 1. *Suppose $\mathcal{F}$ is recourse invariant, and let $f \in \arg\min_{f\in\mathcal{F}} R_P(f)$ be a minimizer of the risk without recourse. Let $f' \in \mathcal{F}$ be the (unique) classifier such that $f'(\varphi_{f'}(x_0)) = f(x_0)$ for all $x_0 \in \mathcal{X}$ and define Δ to be,*

$$\Delta := \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0),Y)] - \mathop{\mathbb{E}}_{(X_0,Y)\sim Q_{f'}}[\ell(f(X_0),Y)]\,.$$

Here, the function f' compensates the effect of giving recourse for the original classifier f, and it exists by recourse invariance. The quantity Δ measures the change in risk when we fix the classifier to be f, but the data are either generated by P (no recourse) or $Q_{f'}$ (recourse for the classifier f'). Intuitively, Δ measures the effect of recourse on the distribution of users when the strategy is to choose a function f' that compensates for the effect of recourse. We generally expect recourse to move users further into the positive class, and therefore to make it more certain that their class label will indeed be $Y = +1$, which means that Δ would be positive. A detailed example is provided in Appendix D.

Theorem 5 (Strategizing in the Compliant Case). *Let ℓ be the $0/1$ loss, assume (1), (2), (3) from Section 2.2 and suppose $\mathcal{F}$ is invariant under recourse. Let Δ be as in Definition 1. Then, the risk after providing recourse in the compliant case can be bounded in terms of the risk without recourse when the party deploying the classifier strategizes to minimize their risk over $\mathcal{F}$,*

$$\min_{f\in\mathcal{F}} R_{Q_f}(f) \le R_{Q_{f'}}(f') = \min_{f\in\mathcal{F}} R_P(f) - \Delta,$$

where f' is as in Definition 1.

When Δ is positive, this shows that providing recourse will be beneficial. In Appendix D, we prove Theorem 5 and

Table 1: Estimated risks on the Census dataset. Lower risk in bold.

	Wachter		GS		CoGS	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
LR	**0.21 ± 0.03**	0.30 ± 0.02	**0.22 ± 0.02**	0.33 ± 0.03	**0.21 ± 0.02**	0.35 ± 0.03
GBT	0.15 ± 0.01	**0.05 ± 0.01**	**0.15 ± 0.02**	**0.17 ± 0.12**	**0.16 ± 0.02**	**0.35 ± 0.25**
DT	**0.25 ± 0.04**	**0.23 ± 0.05**	**0.24 ± 0.04**	**0.46 ± 0.19**	**0.24 ± 0.06**	0.46 ± 0.11
NB	**0.18 ± 0.02**	0.76 ± 0.02	**0.18 ± 0.02**	0.77 ± 0.03	**0.18 ± 0.03**	0.81 ± 0.03
QDA	**0.21 ± 0.02**	0.76 ± 0.03	**0.20 ± 0.02**	0.75 ± 0.04	**0.20 ± 0.02**	0.81 ± 0.03
NN 1	**0.16 ± 0.01**	**0.27 ± 0.11**	**0.16 ± 0.02**	0.25 ± 0.06	**0.16 ± 0.02**	0.28 ± 0.06
NN 2	**0.16 ± 0.02**	0.32 ± 0.05	**0.15 ± 0.02**	0.31 ± 0.04	**0.15 ± 0.02**	0.35 ± 0.06
NN 3	**0.16 ± 0.02**	0.36 ± 0.07	**0.16 ± 0.02**	0.30 ± 0.06	**0.16 ± 0.02**	0.33 ± 0.06
NN 4	**0.16 ± 0.01**	0.38 ± 0.06	**0.16 ± 0.02**	0.34 ± 0.06	**0.15 ± 0.02**	0.38 ± 0.08
NN 5	**0.16 ± 0.01**	0.38 ± 0.06	**0.16 ± 0.02**	0.34 ± 0.05	**0.15 ± 0.02**	0.38 ± 0.08

expand the example of Section 3.1 by showing that $\Delta > 0$ in that case.

5.4 Discussion

We observe that both in the defiant and in the compliant case, an appealing strategy for the party deploying the classifier is to compensate for the effect of recourse by changing their classifier in a way that maintains the original decision boundary. This implies that all users get classified exactly the same way as without recourse, and the only effect of recourse is to change the conditional distribution of Y. For instance, in a bank loan setting, the same customers would get the loan, but some customers might be required to reduce their probability of defaulting before getting it.

6 Experiments

In addition to our theoretical results, we perform several experiments that showcase the possible increase in risk by providing recourse. We conduct these on synthetic data and real data. In both cases we generate Y according to the compliant setting, and except if it is stated otherwise, recourse is provided for all x_0 that are classified as class -1. Further details for all the experiments are available in Appendix E.

6.1 Synthetic Data

The synthetic data consist of the 3 datasets shown in Figure 3, all in 2 dimensions: a Moons dataset, which consists of two translated semi-circles with Gaussian noise; a Circles dataset, which consists of two nested circles with Gaussian noise; and a final dataset consisting of 2 Gaussians with different means and covariances. Counterfactuals for $c(x_0, z) = \|z - x_0\|$ were computed by a brute force search to find the closest point z with $f(z) = +1$ from a dense grid over $\mathcal{X}$.

A summary of the estimated risks for a variety of classifiers can be seen in Table 2. The experiments were also repeated 10 times to estimate confidence bounds for the risks. We also performed experiments where not everyone accepts the counterfactual. We distinguish between two cases. In the first, everyone has the same probability $r(x_0) = p$ of accepting the counterfactual. In the second case, the probability of accepting is determined by the distance towards the counterfactual explanation. We choose to model this probability as $r(x_0) = e^{-\frac{1}{2\sigma^2}\|x_0 - \varphi(x_0)\|^2}$. These results are summarised in the plots in Figure 4 and show that giving more points recourse, generally increases the risk. The plots show the risk difference $R_Q - R_P$ on the y-axis, and either p or σ on the x-axis. We see a clear linear dependence on p in the first case, which is predicted by our results. See Appendix E.1.1 for a derivation of this fact.

Looking at Table 2, We observe that the risk increases in all cases for the Moons and Gaussians dataset. With the Circles dataset, most of the risks with recourse had a higher mean, but the confidence bounds were overlapping. The biggest exception was logistic regression on the Circles dataset. Here, the risk decrease happens because logistic regression has a linear decision boundary, which is severely inappropriate for this data. Without recourse, almost half of the class $+1$ is misclassified, because the linear boundary cuts both circles. If the points of the outer circle, which are of class -1, are projected onto this line, a large portion will land inside the inner circle, where the conditional probability of class $+1$ will be significantly larger than $\frac{1}{2}$.

6.2 Real Data

For the real datasets we use the *Give me Credit*, *Census Income*, and *Home Equity Line of Credit (HELOC)* datasets, from the `CARLA` Python package (Pawelczyk et al., 2021). All features were normalized to $[0, 1]$. We compare various classifiers, and 3 counterfactual methods: Wachter's method (Wachter et al., 2017), the Growing Spheres method (Laugel et al., 2018), and Counterfactual Genetic Search (CoGS) (Virgolin and Fracaros, 2023). The main challenge on real data is that we do not have access to the true conditional distribution, $P(Y \mid X)$. This distribution is needed to sample Y after obtaining $X = \varphi(X_0)$ through recourse. To circumvent this issue we reserve a large portion of the data to train a calibrated classifier for the conditional probabilities.

A summary of the estimated risks can be seen in Table 1, which is shown here, and Tables 3 and 4 in Appendix E. Every experiment was repeated 10 times to estimate confidence bounds. For the Census and HELOC datasets, providing recourse indeed increases the risk in most cases, as predicted by our theoretical results, but in a small number of cases the risk goes down. The exceptions may be explained if the accuracy of the classifiers in class -1 is not significantly better than random guessing, as required by (9). The third dataset is the Credit data: in this case the classes are very unbalanced (class -1 occurs only 7% of the time), and, because we did not fully finetune the classifiers, several of them perform worse than the 7% base error rate that can be obtained by always predicting class $+1$. Since our theoretical results only apply to high accuracy classifiers (see again Equation 9), this leaves room for the possibility that providing recourse might actually decrease the risk in this case. The estimates in Table 3 indeed suggest that this

Table 2: Estimated risks on synthetic data sets. Lower risk is bold.

	Moons data		Circles data		Gaussians data	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
Logistic Regression (LR)	**0.13 ± 0.01**	0.32 ± 0.03	0.52 ± 0.02	**0.36 ± 0.02**	**0.14 ± 0.01**	0.35 ± 0.05
GradientBoostedTrees (GBT)	**0.07 ± 0.03**	0.25 ± 0.09	**0.18 ± 0.02**	0.27 ± 0.03	**0.14 ± 0.02**	0.35 ± 0.06
Decision Tree (DT)	**0.09 ± 0.02**	0.28 ± 0.08	**0.19 ± 0.02**	0.26 ± 0.04	**0.14 ± 0.02**	0.39 ± 0.08
Naive Bayes (NB)	**0.13 ± 0.02**	0.33 ± 0.03	**0.16 ± 0.02**	**0.16 ± 0.03**	**0.16 ± 0.03**	0.29 ± 0.03
QuadraticDiscriminantAnalysis (QDA)	**0.13 ± 0.01**	0.33 ± 0.04	**0.16 ± 0.01**	**0.16 ± 0.03**	**0.13 ± 0.02**	0.38 ± 0.05
Neural Network(4) (NN 1)	**0.12 ± 0.04**	0.27 ± 0.09	**0.26 ± 0.22**	**0.27 ± 0.09**	**0.14 ± 0.02**	0.37 ± 0.04
Neural Network(4, 4) (NN 2)	**0.07 ± 0.07**	0.25 ± 0.06	**0.18 ± 0.01**	**0.23 ± 0.04**	**0.12 ± 0.00**	0.40 ± 0.00
Neural Network(8) (NN 3)	**0.07 ± 0.06**	0.22 ± 0.04	**0.17 ± 0.02**	**0.20 ± 0.02**	**0.13 ± 0.02**	0.36 ± 0.04
Neural Network(8, 16) (NN 4)	**0.03 ± 0.01**	0.26 ± 0.03	**0.17 ± 0.01**	**0.19 ± 0.04**	**0.13 ± 0.02**	0.39 ± 0.04
Neural Netowrk(8, 16, 8) (NN 5)	**0.03 ± 0.01**	0.26 ± 0.03	**0.17 ± 0.01**	**0.19 ± 0.04**	**0.13 ± 0.02**	0.39 ± 0.04

is happening, but, unfortunately, because of the class imbalance, the variance in our risk estimates is too large to definitively confirm that this indeed occurs.

7 Conclusion

We demonstrated, analytically and empirically, that in many cases the risk will increase when recourse is provided. This implies that recourse can be harmful at the population level, and therefore for a large group of users. In such cases, alternative types of explanations might be called for. One interesting alternative direction is the existing work on contestability, which addresses the question of whether an algorithmic decision is correct according to common sense, moral or legal standards (Freiesleben, 2022). As a possibility for future work, our framework might be extended by also accounting for the cost incurred by the users when implementing recourse, e.g., by adding a scaled version of $c(X_0, X)$ to $R_Q(f)$. Assuming positive costs, this would make recourse even less appealing, and lead to the conclusion that it is harmful in an even larger number of cases. Another extension, which would be more interesting to explore, would be to apply our framework in cases where the users and the party deploying the classifier have different loss functions. Then the relation between f and Bayes-optimal decisions for the user's loss would be broken, which might lead to different conclusions.

Acknowledgments

D. Garreau acknowledges support from the French government, through the NIM-ML project (ANR-21-CE23-0005-01). This work was realized while he was working at Université Côte d'Azur (France). T. Van Erven was supported by the Netherlands Organization for Scientific Research (NWO) under grant number VI.Vidi.192.095.

References

André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *IEEE Symposium Series on Computational Intelligence*, 2021.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR, 2022.

Ricardo Dominguez-Olmedo, Amir H. Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 2022.

Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 2022.

Timo Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32(1):77–109, 2022.

Ruijiang Gao and Lakkaraju Himabindu. On the impact of algorithmic recourse on social segregation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 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*, 2016.

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.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. DACE: distribution-aware counterfactual explanation by mixed-integer linear optimization. In

International Joint Conference on Artificial Intelligence, IJCAI, 2020.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 2022.

Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. A causal perspective on meaningful and robust algorithmic recourse. *arXiv preprint arXiv:2107.07853*, 2021.

Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. Improvement-focused causal recourse (icr). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *Internation Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU. Theory and Foundations*, Communications in Computer and Information Science. Springer, 2018.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arxiv:1906.04774*, 2019.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR, 2018.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*, 2019.

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, PMLR, 2023.

Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. *International Transactions in operational research*, 9(5): 583–597, 2002.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Axel Parmentier and Thibaut Vidal. Optimal counterfactual explanations in tree ensembles. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2021.

Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, PMLR, 2022a.

Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. *arXiv preprint arXiv:2203.06768*, 2022b.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems*, 2020.

Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*, 2019.

Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316, 2023.

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

Jialu Wang Yatong Chen. Linear classifiers that encourage constructive adaptation. *Algorithmic Recourse workshop at ICML'21*, 2021.

A Proofs for Section 2

Recall that, for any $x_0 \in \mathcal{X}$, we choose

$$\varphi(x_0) \in \operatorname*{arg\,min}_{z \in \mathcal{X}:\ f(z)=+1} c(x_0, z)\,.$$

Lemma 6. *Assume that $\{z \in \mathcal{X} \mid f(z) = +1\}$ is a closed subset of $\mathbb{R}^d$, and that the cost c satisfies Assumption (3). Then $\varphi(x_0)$ belongs to the boundary of $\{z \in \mathcal{X} \mid f(z) = +1\}$ for all $x_0 \in \mathcal{X}$ such that $f(x_0) = -1$ and $\varphi(x_0) = x_0$ if x_0 belongs to $\{z \in \mathcal{X} \mid f(z) = +1\}$.*

Proof. First, we remark that $c(x_0, z)$ is minimized by $z = x_0$, whenever $x_0 \in \{z \in \mathcal{X} \mid f(z) = +1\}$, which shows that $\varphi(x_0) = x_0$ for all x_0 in $\{z \in \mathcal{X} \mid f(z) = +1\}$.

Moving towards the case of x_0 such that $f(x_0) = -1$. Let us set $x_1 = \varphi(x_0)$ and $A = \{z \in \mathcal{X} \mid f(z) = +1\}$. By contradiction, assume that x_1 does not belong to the boundary of A. Since by construction $x_1 \in A$, then we must have $x_1 \in A^{\circ}$, the interior of A. Consider the point x_2 that lies on the line segment $[x_0, x]$ and on the boundary of A. As $c(x_0, z)$ is increasing for $z \in [x_0, x]$ it must be that

$$c(x_0, x_2) < c(x_0, x).$$

This contradicts the definition of x_1 and concludes the proof. □

Note, that the requirement on c can be further weakened by assuming that there is a path between x_0 and x such that $z \mapsto c(x_0, z)$ is increasing along this path. So, we do not necessarily need straight line segments. The proof of this statement is analogous to the proof of Lemma 6.

B Further Details for Section 3

Before we start the proof of the main result in Section 3, we will introduce notation and some additional results. The conditional distributions of Y given X_0 will be denoted by

$$\begin{aligned} p_+(x) &:= P(Y = +1 \mid X_0 = x) \\ &= 1 - P(Y = -1 \mid X_0 = x) =: 1 - p_-(x). \end{aligned}$$

Now, we will prove a general result about expressing the risk under Q, of which Theorem 1 is a consequence. Every expectation $\mathbb{E}$ will be with respect to P in this section.

Lemma 7. *Let ℓ be a loss function with $\ell(y, y) = 0$, and assume the setting of Section 2.2 (i.e., (1), (2), (3)). Suppose that $P(Y = 1|X_0 = x) = \frac{1}{2}$ for all x on the decision boundary of f. Then*

(a) For the defiant case,

$$\begin{aligned} R_Q(f) = \ell(1,-1)P(Y = -1)\mathbb{E}\left[r(X_0)|Y = -1\right] \\ + \mathbb{E}\left[(1 - r(X_0))\ell(f(X_0), Y)\right]; \end{aligned} \tag{11}$$

(b) For the compliant case,

$$R_Q(f) = \ell(1,-1)\mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right] \tag{12}$$
$$+ \mathbb{E}\left[(1 - r(X_0))\ell(f(X_0), Y)\right]. \tag{13}$$

Proof. Before distinguishing between the 2 cases, we expand the expression for the risk under Q as

$$\begin{aligned} R_Q(f) &= \int_{\mathcal{X}^2 \times \mathcal{Y}} \ell(f(x), y)Q(\mathrm{d}y \mid x, x_0)Q(\mathrm{d}x|x_0)P(\mathrm{d}x_0) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} r(x_0)\ell(f(\varphi(x_0)), y)Q(\mathrm{d}y \mid \varphi(x_0), x_0)P(\mathrm{d}x_0) \\ &\quad + \int_{\mathcal{X} \times \mathcal{Y}} (1 - r(x_0))\ell(f(x_0), y)P(\mathrm{d}y \mid x_0)P(\mathrm{d}x_0) \\ &= \ell(1,-1)\int_{\mathcal{X}} r(x_0)Q(Y = -1 \mid \varphi(x_0), x_0)P(\mathrm{d}x_0) \\ &\qquad\qquad\qquad\qquad (\text{since } f(\varphi(x_0)) = 1) \\ &\quad + \mathbb{E}\left[(1 - r(X_0))\ell(f(X_0), Y)\right]. \end{aligned}$$

We focus now on the integral in the first term.

Defiant Case: The first term in the expression in the above display becomes

$$\begin{aligned} &\int_{\mathcal{X}} r(x_0)Q(Y = -1 \mid \varphi(x_0), x_0)P(\mathrm{d}x_0) \\ &= \int_{\mathcal{X}} r(x_0)P(Y = -1 \mid x_0)P(\mathrm{d}x_0) \\ &= P(Y = -1)\int_{\mathcal{X}} r(x_0)P(\mathrm{d}x_0 \mid Y = -1) \\ &= P(Y = -1)\mathbb{E}\left[r(X_0) \mid Y = -1\right]. \end{aligned}$$

Compliant Case: The first term now becomes

$$\begin{aligned}
&\int_{\mathcal{X}} r(x_0)Q(\mathrm{d}y \mid \varphi(x_0), x_0)P(\mathrm{d}x_0) \\
&= \int_{\mathcal{X}} r(x_0)P(\mathrm{d}y \mid \varphi(x_0))P(\mathrm{d}x_0) \\
&= \int_{\mathcal{X}} r(x_0)P(Y=-1 \mid \varphi(x_0))P(\mathrm{d}x_0) \\
&= \int_{\mathcal{X}} r(x_0)p_-(\varphi(x_0))P(\mathrm{d}x_0) \\
&= \mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right].
\end{aligned}$$

□

We are now ready to prove Theorem 1.

Theorem 1 (Bayes-Optimal Classifier Risk Increase). *Let ℓ be the $0/1$ loss, and assume the setting of Section 2.2 (i.e., (1), (2), (3)). Suppose that $P(Y=1|X_0=x)=\frac{1}{2}$ for all x on the decision boundary of f_P^*. Then*

(a) For the defiant case,

$$\begin{aligned}
R_Q(f_P^*) &= P(B=1, f_P^*(X_0)=-1, Y=-1) \\
&\quad - P(B=1, f_P^*(X_0)=-1, Y=+1) \quad (4)\\
&\quad + R_P(f_P^*) \\
&\geq R_P(f_P^*);
\end{aligned}$$

(b) For the compliant case,

$$\begin{aligned}
R_Q(f_P^*) &= \tfrac{1}{2}P(B=1, f_P^*(X_0)=-1) \\
&\quad - P(B=1, f_P^*(X_0)=-1, Y=1) \quad (5)\\
&\quad + R_P(f_P^*) \\
&\geq R_P(f_P^*).
\end{aligned}$$

Both inequalities are strict if $P(B=1, f_P^*(X_0)=-1) > 0$, i.e., if the probability of recourse in the negative class is non-zero.

Proof. For both cases, we will first prove the equality and then show that the expectation is always non-negative for the inequality. From those proofs it can be seen how the strict inequality is derived. We apply Lemma 7 to both cases. Remark that $\ell(1,-1)=\ell(-1,1)=1$ and rewrite the common term as

$$\begin{aligned}
&\mathbb{E}\left[(1-r(X_0))\ell(f_P^*(X_0), Y)\right] \\
&= R_P(f_P^*) - \mathbb{E}\left[r(X_0)\mathbb{1}\left(f_P^*(X_0)\neq Y\right)\right] \\
&= R_P(f_P^*) - P(B=1, f_P^*(X_0)\neq Y). \quad (14)
\end{aligned}$$

Defiant Case: In this case, we rewrite the first term in (11) to get

$$P(Y=-1)\mathbb{E}\left[r(X_0)|Y=-1\right] = \mathbb{E}\left[r(X_0)\mathbb{1}(Y=-1)\right]. \quad (15)$$

Combining expressions (14) and (15) gives the result,

$$\begin{aligned}
R_Q(f_P^*) &= \mathbb{E}\left[r(X_0)(\mathbb{1}(Y=-1) - \mathbb{1}(f_P^*(X_0)\neq Y))\right] \\
&\quad + R_P(f_P^*) \quad (16)\\
&= P(B=1, Y=-1) - P(B=1, f_P^*(X_0)\neq Y) \\
&\quad + R_P(f_P^*) \\
&= P(B=1, f_P^*(X_0)=-1, Y=-1) \\
&\quad - P(B=1, f_P^*(X_0)=-1, Y=+1) \\
&\quad + R_P(f_P^*).
\end{aligned}$$

It remains to show that the difference of the first two probabilities is positive. We return to the formulation in terms of expectations and indicator functions. We can rewrite the indicator functions giving rise to those probabilities as

$$\begin{aligned}
&\mathbb{1}(f_P^*(x_0)=-1, y=-1) - \mathbb{1}(f_P^*(x_0)=-1, y=1) \\
&= \mathbb{1}(f_P^*(x_0)=-1)(\mathbb{1}(y=-1) - \mathbb{1}(y=1)).
\end{aligned}$$

The expectation in (16) now becomes

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} r(x_0)(\mathbb{1}(y=-1) - \mathbb{1}(f_P^*(x_0)\neq y))P(\mathrm{d}x_0, \mathrm{d}y) \\
&= \int_{\{f_P^*=-1\}\times\mathcal{Y}} r(x_0)(\mathbb{1}(y=-1) \quad (17)\\
&\qquad - \mathbb{1}(y=1))P(\mathrm{d}y \mid X=x)P(\mathrm{d}x_0) \\
&= \int_{\{f_P^*=-1\}} r(x_0)(p_-(x_0) - p_+(x_0))P(\mathrm{d}x_0). \quad (18)
\end{aligned}$$

Now, by f_P^* being the Bayes optimal classifier we know that $p_-(x_0)\geq p_+(x_0)$ for all x_0 on $\{f_P^*=-1\}$. So, we see that the integral in (17) is non-negative.

Compliant Case: We note that $p_+(\varphi(x_0)) = p_-(\varphi(x_0)) = \frac{1}{2}$ for any x_0 with $f_P^*(x_0)=-1$, because those points are projected onto the decision boundary by assumption (3) and Lemma 6. The points on the decision boundary of the Bayes classifier are exactly where the probability of being either class is $\frac{1}{2}$, by assumption. The first expectation in (12) can now be written as

$$\begin{aligned}
\mathbb{E}[r(X_0)p_-(\varphi(X_0))] &= \tfrac{1}{2}\mathbb{E}[r(X_0)\mathbb{1}(f_P^*(X_0)=-1)] \\
&\quad + \mathbb{E}[r(X_0)p_-(X_0)\mathbb{1}(f_P^*(X_0)=+1)] \\
&= \tfrac{1}{2}P(B=1, f_P^*(X_0)=-1) \\
&\quad + P(B=1, f_P^*(X_0)=+1, Y=-1). \quad (19)
\end{aligned}$$

We note that the second probability in (19) cancels the second probability (14) partly. First we write the latter probability as

$$\begin{aligned}
&P(B=1, f_P^*(X_0)\neq Y) \\
&= P(B=1, f_P^*(X_0)=1, Y=-1) \\
&\quad + P(B=1, f_P^*(X_0)=-1, Y=1).
\end{aligned}$$

Subtracting both probabilities gives

$$\begin{aligned}P(B=1,&f_P^*(X_0)=1,Y=-1)-P(B=1,f_P^*(X_0)\neq Y)\\&=-P(B=1,f_P^*(X_0)=-1,Y=1).\end{aligned}$$

Substituting (19) and (14) into the expression for $R_Q(f_P^*)$ gives

$$\begin{aligned}R_Q(f_P^*)&=\mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right]\\&\quad+\mathbb{E}\left[(1-r(X_0))\ell(f(X_0),Y)\right]\\&=\mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right]+R_P(f_P^*)\\&\quad-P(B=1,f_P^*(X_0)\neq Y)\\&=\tfrac{1}{2}P(B=1,f_P^*(X_0)=-1)\\&\quad-P(B=1,f_P^*(X_0)=-1,Y=1)+R_P(f_P^*).\end{aligned}$$

To derive the necessary inequality, we focus again on the first two probabilities and write this explicitly as an integral. This integral is given by

$$\begin{aligned}\tfrac{1}{2}P(B=1,&f_P^*(X_0)=-1)\\&-P(B=1,f_P^*(X_0)=-1,Y=1)\\&=\int_{\{f_P^*=-1\}}r(x_0)(\tfrac{1}{2}-p_+(x_0))P(\mathrm{d}x_0)\\&\geq 0\end{aligned}$$

Where we have used that the on the set $\{x_0\in\mathcal{X}\mid f_P^*(x_0)=-1\}$ it must be that $p_+(x_0)\leq\frac{1}{2}$, because f_P^* is the Bayes classifier.

The strict inequality follows by remarking that the difference of the integrand in both integrals of the defiant and compliant case will be strictly positive on some positive probability set, if $P(B=1,f_P^*=-1)>0$. $\square$

B.1 Additional Details Gaussian Example in Section 3.1

In Section 3.1 it is claimed that the Bayes risk can be written as $R_P(f_P^*)=\Phi(-\frac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}})$. Here, we show this and additionally derive the Bayes optimal classifier for general $\mu,\nu\in\mathbb{R}^d$.

The conditional distribution can be calculated explicitly. Let $p_\mu(x)=e^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)}$ and $p_\nu(x)=e^{-\frac{1}{2}(x-\nu)^\top\Sigma^{-1}(x-\nu)}$, then

$$\begin{aligned}P(Y=1&\mid X_0=x)\\&=\frac{P(Y=1)p_\mu(x)}{P(Y=1)p_\mu(x)+P(Y=-1)p_\nu(x)}\\&=\frac{1}{1+e^{-x^\top\Sigma^{-1}(\mu-\nu)+\frac{1}{2}(\|\mu\|^2_{\Sigma^{-1}}-\|\nu\|^2_{\Sigma^{-1}})}}.\end{aligned}$$

From this we see that for $\theta=\Sigma^{-1}(\mu-\nu)$ and $\theta_0=-\frac{1}{2}(\|\mu\|^2_{\Sigma^{-1}}-\|\nu\|^2_{\Sigma^{-1}})$ the Bayes classifier is given by $f_P^*(x)=\mathrm{sign}(x^\top\theta+\theta_0)$. We can now calculate the Bayes risk by first rewriting this risk as

$$\begin{aligned}R_P(f_P^*)&=\tfrac{1}{2}P(f_P^*(X_0)=-1\mid Y=1)\\&\quad+\tfrac{1}{2}P(f_P^*(X_0)=1\mid Y=-1)\\&=\tfrac{1}{2}P(X_0^\top\theta+\theta_0<0\mid Y=1)\\&\quad+\tfrac{1}{2}P(X_0^\top\theta+\theta_0\geq 0\mid Y=-1).\end{aligned}\tag{20}$$

As X_0 is Gaussian, conditional on Y, we know that $X_0^\top\theta+\theta_0$ is also Gaussian. For $Y=1$, we get $\mathcal{N}(\mu^\top\theta+\theta_0,\|\theta\|^2_{\Sigma^{-1}})$ and for $Y=-1$ we get $\mathcal{N}(\nu^\top\theta+\theta_0,\|\theta\|^2_{\Sigma^{-1}})$. Translating and rescaling allows us to rewrite the probabilities in (20) in terms of the CDF Φ of the standard normal distribution,

$$\begin{aligned}P(X_0^\top\theta+\theta_0<0\mid Y=1)&=\Phi\left(\frac{-\mu^\top\theta-\theta_0}{\|\theta\|_{\Sigma^{-1}}}\right)\\=\Phi\left(\frac{-\|\mu\|^2_{\Sigma^{-1}}+\mu^\top\Sigma^{-1}\nu+\frac{1}{2}(\|\mu\|^2_{\Sigma^{-1}}-\|\nu\|^2_{\Sigma^{-1}})}{\|\mu-\nu\|_{\Sigma^{-1}}}\right)\\=\Phi\left(\frac{-\frac{1}{2}\|\mu-\nu\|^2_{\Sigma^{-1}}}{\|\mu-\nu\|_{\Sigma^{-1}}}\right)&=\Phi(-\tfrac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}}).\end{aligned}$$

Analogously, we would get

$$P(X_0^\top\theta+\theta_0\geq 0\mid Y=-1)=\Phi(-\tfrac{1}{2}\|\mu-\nu\|_{\Sigma^{-1}}).$$

Combining the two probabilities gives the desired result.

C Proofs of Section 4

In this section we present all the previous and additional results of Section 4.

C.1 Proof of Theorem 2

Theorem 2 (Probabilistic Classifier Risk Increase, $0/1$ loss). *Let ℓ be the $0/1$ loss. Let $g:\mathcal{X}\to[0,1]$ be a continuous, probabilistic classifier, and define $f(x)=\mathrm{sign}(g(x)-\frac{1}{2})$. Assume (1), (2), (3) from Section 2.2. Then,*

(a) For the defiant case,

$$\begin{aligned}R_Q(f)&=P(B=1,f(X_0)=-1,Y=-1)\\&\quad-P(B=1,f(X_0)=-1,Y=+1)\\&\quad+R_P(f).\end{aligned}\tag{6}$$

Moreover, $R_Q(f)\geq R_P(f)$ if and only if

$$P(Y=-1\mid B=1,f(X_0)=-1)\geq\tfrac{1}{2}.\tag{7}$$

If we additionally assume that g satisfies (B) with $0\leq\varepsilon\leq\frac{1}{2}$, then

(b) For the compliant case, $R_Q(f)$ is lower and upper bounded by

$$\begin{aligned}(\tfrac{1}{2}\pm\varepsilon)&P(B=1,f(X_0)=-1)\\&+P(f(X_0)=+1,Y=-1)\\&+P(B=0,f(X_0)=-1,Y=1).\end{aligned}\tag{8}$$

Moreover, $R_Q(f) \geq R_P(f)$ *if*

$$P(Y = -1 \mid B = 1, f(X_0) = -1) \geq \tfrac{1}{2} + \varepsilon. \quad (9)$$

Proof. **Defiant Case:** We again use Lemma 7 which gives us

$$\begin{aligned} R_Q(f) &= P(Y = -1)\,\mathbb{E}[r(X_0) \mid Y = -1] \\ &\quad + \mathbb{E}[(1 - r(X_0)\mathbb{1}(f(X_0) \neq Y)] \\ &= P(Y = -1, B = 1) + P(f(X_0) \neq Y) \\ &\quad - P(f(X_0) \neq Y, B = 1) \\ &= P(Y = -1, B = 1) - P(f(X_0) \neq Y, B = 1) \\ &\quad + R_P(f) \\ &= P(B = 1, f(X_0) = -1, Y = -1) \\ &\quad - P(B = 1, f(X_0) = -1, Y = +1) \\ &\quad + R_P(f). \end{aligned}$$

To derive the second claim, we upper bound $R_P(f)$ by $R_Q(f)$. We see that the $R_P(f)$ term drops on both sides and we are left with

$$\begin{aligned} P(B = 1, &f(X_0) = -1, Y = -1) \leq \\ &P(B = 1, f(X_0) = -1, Y = +1) \end{aligned}$$

Conditioning on $\{B = 1, f(X_0) = -1\}$ and cancelling the common terms gives us

$$\begin{aligned} &P(Y = +1 \mid f(X_0) = -1, B = 1) \\ &\quad \leq P(Y = -1 \mid f(X_0) = -1, B = 1), \qquad \iff \\ &P(Y = -1 \mid f(X_0) = -1, B = 1) \geq \tfrac{1}{2}. \end{aligned}$$

Compliant Case: We apply Lemma 7. Note, that Assumption B and Lemma 6 tell us that on the set $\{x_0 \in \mathcal{X} \mid f(x_0) = -1\}$ we have that $\frac{1}{2} - \varepsilon < p_-(\varphi(X_0)) \leq \frac{1}{2} + \varepsilon$ in expectation. For the first expectation we get the upper bound

$$\begin{aligned} &\mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right] \\ &\quad = \mathbb{E}\left[r(X_0)\mathbb{1}\{f(X_0) = -1\}p_-(\varphi(X_0))\right] \\ &\qquad + \mathbb{E}\left[r(X_0)\mathbb{1}\{f(X_0) = 1\}p_-(X_0)\right] \\ &\quad \leq (\tfrac{1}{2} + \varepsilon)P(f(X_0) = -1, B = 1) \\ &\qquad + P(f(X_0) = 1, Y = -1, B = 1). \end{aligned}$$

Analogously, for the lower bound we get

$$\begin{aligned} \mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right] \geq (\tfrac{1}{2} - \varepsilon)&P(f(X_0) = -1, B = 1) \\ + &P(f(X_0) = 1, Y = -1, B = 1). \end{aligned}$$

We write the second expectation as follows in this case,

$$\mathbb{E}\left[(1 - r(X_0))\ell(f(X_0), Y)\right] = P(f(X_0) \neq Y, B = 0).$$

This leaves us with

$$\begin{aligned} R_Q(f) &\leq (\tfrac{1}{2} + \varepsilon)P(f(X_0) = -1, B = 1) \\ &\qquad + P(f(X_0) = 1, Y = -1, B = 1) \\ &\qquad + P(B = 0, f(X_0) \neq Y) \\ &\leq (\tfrac{1}{2} + \varepsilon)P(f(X_0) = -1, B = 1) \\ &\qquad + P(f(X_0) = 1, Y = -1) \\ &\qquad + P(f(X_0) = -1, Y = 1, B = 0). \end{aligned} \quad (21)$$

Similarly, for the lower bound we get

$$\begin{aligned} R_Q(f) \geq (\tfrac{1}{2} - \varepsilon)&P(f(X_0) = -1, B = 1) \\ &+ P(f(X_0) = 1, Y = -1) \\ &+ P(f(X_0) = -1, Y = 1, B = 0) \end{aligned} \quad (22)$$

Combining expressions (21) and (22) gives the desired lower and upper bound.

We move to the second claim. This time, we upper bound $R_P(f)$ by the derived lower bound. This gives us

$$\begin{aligned} &P(Y = 1, f(X_0) = -1) \\ &\quad \leq (\tfrac{1}{2} - \varepsilon)P(f(X_0) = -1, B = 1) \\ &\qquad + P(f(X_0) = -1, Y = 1, B = 0) \qquad \iff \\ &P(Y = 1, f(X_0) = -1, B = 1) \\ &\quad \leq (\tfrac{1}{2} - \varepsilon)P(f(X_0) = -1, B = 1) \qquad \iff \\ &P(Y = 1 \mid f(X_0) = -1, B = 1) \\ &\quad \leq (\tfrac{1}{2} - \varepsilon)P(Y = -1 \mid f(X_0) = -1, B = 1) \end{aligned}$$

The final inequality can be rewritten as

$$P(Y = -1 \mid f(X_0) = -1, B = 1) \geq (\tfrac{1}{2} + \varepsilon).$$

□

C.2 Proof of Theorem 3

Theorem 3 (Probabilistic Classifier Risk Increase, Surrogate Loss). *Let* $\ell : [0, 1] \times \{-1, +1\} \to \mathbb{R}$ *be any loss such that* $\ell(1/2, -1) = \ell(1/2, +1) = c$ *for some constant* c*. Let* $g : \mathcal{X} \to [0, 1]$ *be a continuous, probabilistic classifier, and define* $f(x) = \operatorname{sign}(g(x) - \frac{1}{2})$*. Further assume* (1), (2), (3) *from Section* 2.2*. Then, both for the defiant and for the compliant case, we have* $R_Q(g) \geq R_P(g)$ *if and only if*

$$\mathop{\mathbb{E}}_{P}\left[\ell(g(X_0), Y) \mid f(X_0) = -1, B = 1\right] \leq c. \quad (10)$$

Proof. Let $I = \mathbb{1}\{f(X_0) = -1, B = 1\}$ be the indicator for recourse in the negative class. Then, since $\varphi(X_0)$ lies on the decision boundary (DB) when X_0 is in the negative class,

$$
\begin{aligned}
R_Q(g) &= \mathop{\mathbb{E}}_{(X,Y)\sim Q}[\ell(g(X),Y)] \\
&= \mathop{\mathbb{E}}_{(X_0,Y)\sim Q}[\ell(g(\varphi(X_0)),Y)I] \\
&\quad + \mathop{\mathbb{E}}_{(X,Y)\sim Q}[\ell(g(X),Y)(1-I)] \\
&= \mathop{\mathbb{E}}_{(X_0,Y)\sim Q}[\ell(1/2,Y)I] \qquad (\varphi(X_0) \text{ on the DB}) \\
&\quad + \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(g(X_0),Y)(1-I)] \\
&= cP(f(X_0)=-1,B=1) \quad (\text{by definition of } c) \\
&\quad + \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(g(X_0),Y)(1-I)] \\
&\geq \mathop{\mathbb{E}}_{P}[\ell(g(X_0),Y)I] \\
&\quad + \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(g(X_0),Y)(1-I)] \\
&= R_P(g),
\end{aligned}
$$

where the inequality is equivalent to (10). □

D Additional results and proofs for Section 5

D.1 Examples of classifiers invariant under recourse

Let us justify more rigorously the linear classifier example introduced in Section 5. A visual representation describing this example can be found in Figure 5.

Example 1. Consider the set of linear classifiers $\mathcal{F} = \{f_{\theta,\theta_0}(x) = \operatorname{sign}(x^\top\theta + \theta_0) \mid \theta \in \mathbb{R}^d, \theta_0 \in \mathbb{R}\}$ with the convention that $\operatorname{sign}(z) = +1$ for $z \geq 0$ and $\operatorname{sign}(z) = -1$ otherwise. If the recourse map is such that any point x_0 within distance $D > 0$ of the decision boundary of f_{θ,θ_0} gets mapped to the positive class, then this class is invariant under recourse, because

$$
f_{\theta,\theta_0'}(\varphi(f_{\theta,\theta_0'},x_0)) = f_{\theta,\theta_0}(x_0) \qquad \text{for all } x_0 \in \mathcal{X}
$$

when $\theta_0' = \theta_0 - D\|\theta\|$. To see this, note that the (signed) distance from x_0 to the decision boundary for $f_{\theta,\theta_0'}$ is $\frac{-x_0^\top\theta-\theta_0'}{\|\theta\|}$. Hence the following are all equivalent:

$$
\begin{aligned}
f_{\theta,\theta_0'}(\varphi(f_{\theta,\theta_0'},x_0)) &= +1 \\
\frac{-x_0^\top\theta-\theta_0'}{\|\theta\|} &\leq D \\
x_0^\top\theta + \theta_0' &\geq -D\|\theta\| \\
x_0^\top\theta + \theta_0 &\geq 0 \\
f_{\theta,\theta_0}(x_0) &= +1.
\end{aligned}
$$

One can extend this idea to other geometrical shapes:

Example 2. Consider the spherical classifiers for which $f_{\theta,b}(x) = +1$ if and only if $\|x-\theta\| \geq b$. Then the set $\mathcal{F} = \{f_{\theta,b} \mid \theta \in \mathbb{R}^d, b \in \mathbb{R}_+\}$ is invariant under recourse when the recourse map is again such that any point x_0 in the negative class that lies within distance $D > 0$ of the decision boundary of $f_{\theta,b}$ gets mapped to the positive class. This follows because providing recourse has the effect of effectively shrinking b by D, so we can undo this effect by increasing b to $b' = b + D$:

$$
f_{\theta,b'}(\varphi(f_{\theta,b'},x_0)) = f_{\theta,b}(x_0) \qquad \text{for all } x_0 \in \mathcal{X}.
$$

D.2 Proof of Theorem 4

Theorem 4 (Strategizing in the Defiant Case). *Let ℓ be the $0/1$ loss, assume (1), (2), (3) from Section 2.2 with $r(x_0) \in \{0,1\}$ for all $x_0 \in \mathcal{X}$, and suppose $\mathcal{F}$ is invariant under recourse. Then, providing recourse in the defiant case does not change the risk when the party deploying the classifier strategizes to minimize their risk over $\mathcal{F}$:*

$$
\min_{f\in\mathcal{F}} R_{Q_f}(f) = \min_{f\in\mathcal{F}} R_P(f)\,.
$$

Proof. Note that in the defiant case $Q(Y \mid X_0) = P(Y \mid X_0)$ as

$$
\begin{aligned}
Q(Y \mid X_0) &= \int_{\mathcal{X}} Q(Y, X = \mathrm{d}x \mid X_0) \\
&= \int_{\mathcal{X}} Q(Y \mid X = x, X_0) Q(X = \mathrm{d}x \mid X_0) \\
&= P(Y \mid X_0) \int_{\mathcal{X}} Q(X = \mathrm{d}x \mid X_0) \\
&= P(Y \mid X_0).
\end{aligned}
$$

Using this we, we write

$$
\begin{aligned}
\min_{f\in\mathcal{F}} R_{Q_f}(f) &= \min_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(X,Y)\sim Q_f}[\ell(f(X),Y)] \\
&= \min_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(X_0,Y)\sim Q_f}[\ell(f(\varphi_f(X_0)),Y)] \\
&= \min_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(\varphi_f(X_0)),Y)] \\
&\qquad\qquad (\text{by defiant case}) \\
&= \min_{f\in\mathcal{F}_\varphi^r} \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0),Y)] \\
&\qquad\qquad (\text{by definition of } \mathcal{F}_\varphi^r) \\
&= \min_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0),Y)] \\
&\qquad\qquad (\text{since } \mathcal{F} = \mathcal{F}_\varphi^r) \\
\min_{f\in\mathcal{F}} R_{Q_f}(f) &= \min_{f\in\mathcal{F}} R_P(f)\,.
\end{aligned}
$$

□

D.3 Explicit Δ bound

Example 3. Let us further specialize the setting of Example 1 to the task of distinguishing between two Gaussians with different means $\mu, \nu \in \mathbb{R}^d$ and common positive definite covariance matrix Σ. That is, let $P(Y =$

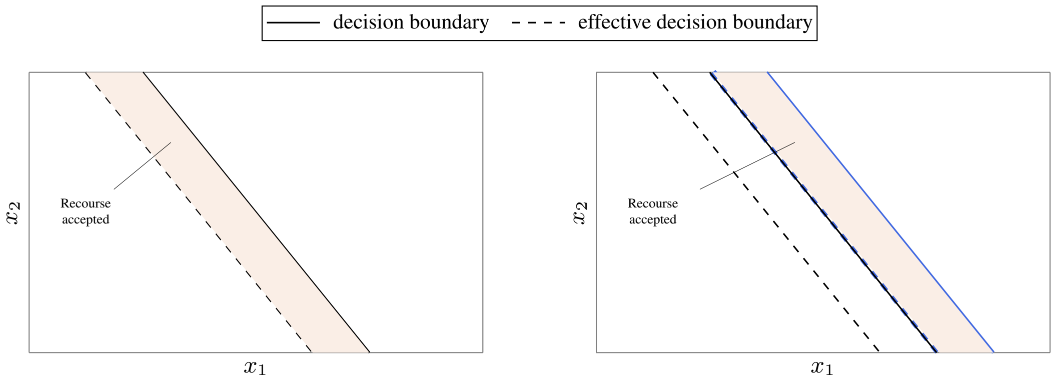

Figure 5: Left figure: Linear classifier with a shaded area to indicate where recourse is accepted. Right figure: The same linear classifier but shifted towards the right in such a way that the effective decision boundary is the original decision boundary.

$-1) = P(Y = +1) = 1/2$, and the data is distributed according to $P(X_0|Y = +1) = \mathcal{N}(\mu, \Sigma)$ and $P(X_0|Y = -1) = \mathcal{N}(\nu, \Sigma)$. Then $f = f_{\theta,\theta_0}$ for $\theta = \Sigma^{-1}(\mu - \nu)$ and $\theta_0 = -\frac{1}{2}(\mu + \nu)^\top \Sigma^{-1}(\mu - \nu)$ is the Bayes optimal classifier. See Section B.1 for a justification. The compensating classifier is given by $f' = f_{\theta,\theta_0'}$ for $\theta_0' = \theta_0 - D\|\theta\|$. Then recourse for f' affects users in a band of width $D\|\theta\|$ which lies just in the positive class according to f:

$$\mathcal{A} = \{x : 0 \leq x^\top \theta + \theta_0 < D\|\theta\|\},$$

and

$$\begin{aligned}\Delta &= \int_{\mathcal{A}} \Big\{P(Y = -1 \mid X_0 = x_0) \\ &\quad - P(Y = -1 \mid X_0 = \varphi_{f'}(x_0))\Big\} P(\mathrm{d}x_0) \\ &= \int_{\mathcal{A}} \Big\{P(Y = -1 \mid X_0 = x_0) \\ &\quad - P\left(Y = -1 \mid X_0 = x_0 - \frac{f'(x_0)}{\|\theta\|^2}\theta\right) \Big\} P(\mathrm{d}x_0).\end{aligned}$$

Again by Section B.1, we can write the posterior distribution as

$$P(Y = -1|X_0 = x) = \frac{1}{1 + e^{\theta^\top x + \theta_0}},$$

Now, we write

$$\begin{aligned}\theta^\top \varphi(f', x_0) + \theta_0 &= \theta^\top x_0 + \theta_0 - (\theta)^\top \left(\frac{f'(x_0)}{\|\theta\|^2}\theta\right) \\ &= \theta^\top x_0 + \theta_0 + 1\,,\end{aligned}$$

since $f'(x_0) = -1$. Thus, $\theta^\top \varphi_{f'}(x_0) + \theta > \theta^\top x_0 + \theta_0$. Since the mapping $t \mapsto 1/(1 + \exp(t))$ is decreasing, we deduce that $\Delta > 0$ in this case.

D.4 Proof of Theorem 5

Theorem 5 (Strategizing in the Compliant Case). *Let ℓ be the $0/1$ loss, assume (1), (2), (3) from Section 2.2 and suppose $\mathcal{F}$ is invariant under recourse. Let Δ be as in Definition 1. Then, the risk after providing recourse in the compliant case can be bounded in terms of the risk without recourse when the party deploying the classifier strategizes to minimize their risk over $\mathcal{F}$,*

$$\min_{f\in\mathcal{F}} R_{Q_f}(f) \leq R_{Q_{f'}}(f') = \min_{f\in\mathcal{F}} R_P(f) - \Delta,$$

where f' is as in Definition 1.

Proof. As in the defiant case, the proof considers the strategic choice f', which compensates for the effect of recourse in order to maintain the same decision boundary as in the case without recourse. The effect of recourse is then only to change the distribution in a way that is captured by Definition 1. We first notice that $\min_{f\in\mathcal{F}} R_{Q_f}(f) \leq R_{Q_{f'}}(f')$ holds because $f' \in \mathcal{F}$. We write

$$\begin{aligned}R_{Q_{f'}}(f') &= \mathop{\mathbb{E}}_{(X,Y)\sim Q_{f'}}[\ell(f'(X), Y))] \\ &= \mathop{\mathbb{E}}_{(X_0,Y)\sim Q_{f'}}[\ell(f'(\varphi_{f'}(X_0)), Y)] \\ &= \mathop{\mathbb{E}}_{(X_0,Y)\sim Q_{f'}}[\ell(f(X_0), Y)] \\ &= \mathop{\mathbb{E}}_{(X_0,Y)\sim P}[\ell(f(X_0), Y)] - \Delta \\ &\qquad\qquad\text{(definition of }\Delta) \\ R_{Q_{f'}}(f') &= \min_{f\in\mathcal{F}} R_P(f) - \Delta\,.\end{aligned}$$

□

E Details Experimental Setup

All the code to reproduce the experiments and figures in this paper can be found in a GitHub repository.[3] All experiments were performed on a single 32-core CPU node (AMD Rome 7H12) with 64GB of RAM. For all experiments we used the

[3] github.com/HiddeFok/consequences-of-recourse

following classifiers from the `scikit-learn` (version 1.0) library:

- `LogisticRegression`, with the default parameters, except for `class_weigt='balanced'`.
- `GradientBoostingClassifier`, with the default parameters, except for `n_estimators=10` .
- `DecisionTreeClassifier`, with the default parameters, except for `class_weigt='balanced'` and `max_depth=4`.
- `GaussianNB`, with the default parameters.
- `RandomForestClassifier`, with the default parameters except for `class_weigt='balanced'`, `max_depth=4` and `n_estimators=10`.
- `QuadraticDiscrimantAnalysis`, with the default parameters.
- `MLPClassifier`, with the default parameters and hidden layers, `hidden_layer_sizes=(4,)`.
- `MLPClassifier`, with the default parameters and hidden layers, `hidden_layer_sizes=(4, 4)`.
- `MLPClassifier`, with the default parameters and hidden layers, `hidden_layer_sizes=(8,)`.
- `MLPClassifier`, with the default parameters and hidden layers, `hidden_layer_sizes=(8, 16)`.
- `MLPClassifier`, with the default parameters and hidden layers, `hidden_layer_sizes=(8, 16, 8)`.

For both the synthetic data experiments and real data experiments, we repeated the experiments 10 times to estimate confidence bounds. The result with the lower risk is written in bold in the result tables. If the confidence bounds overlap we make both results bold.

E.1 Synthetic Data

For each of the synthetic data experiments, we generated 5000 training samples and 1000 test samples with balanced classes, i.e. $P(Y=+1)=P(Y=-1)=\frac{1}{2}$. Examples of the results of these experiments can be found in Figures 6, 7, and 8.

Moons dataset The Moons dataset is acquired through the `make_moons` function from `scikit-learn`. A data point X_0 is sampled by first discretizing $[0,\pi)$ uniformly and drawing one of these points U uniformly, without replacement. Then, sample $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_2)$ and construct X_0 by setting

$$X_0 := s(U)+\varepsilon = \begin{bmatrix}\cos(U)\\ \sin(U)\end{bmatrix}+\varepsilon \qquad \text{for } Y=+1$$
$$X_0 := c(U)+\varepsilon = \begin{bmatrix}1-\cos(U)\\ 1-\sin(U)\end{bmatrix} - \begin{bmatrix}0\\ \frac{1}{2}\end{bmatrix}+\varepsilon \quad \text{for } Y=-1.$$

For our examples we selected $\sigma = 0.2$.

To re-sample the label Y after providing recourse we calculate the conditional distribution of this model. Let, p be the density of ε. Then, the density of $X_0 \mid Y=+1$ and $X_0 \mid Y=-1$, denoted by g_+ and g_- respectively, is given by

$$g_+(x_0) = \frac{1}{\pi}\int_0^\pi p(x_0 - s(u))\mathrm{d}u,$$
$$g_-(x_0) = \frac{1}{\pi}\int_0^\pi p(x_0 - c(u))\mathrm{d}u.$$

In our implementation this integral is approximated by a Riemann sum. The conditional distribution now follows,

$$\begin{aligned} P(Y=1 \mid X_0=x_0) &= \frac{g_+(x_0)P(Y=1)}{g_+(x_0)P(Y=1)+g_-(x_0)P(Y=-1)} \\ &= \frac{g_+(x_0)}{g_+(x_0)+g_-(x_0)}. \end{aligned}$$

Circles dataset The Circles dataset is acquired through the `make_circles` function from `scikit-learn`. A data point X is sampled by first discretizing $[0,2\pi)$ uniformly and drawing one of these points U uniformly, without replacement. Then, sample $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_2)$, set $\lambda \in (0,1)$ and construct X by setting

$$X_0 := \lambda s(U)+\varepsilon = \lambda\begin{bmatrix}\cos(U)\\ \sin(U)\end{bmatrix}+\varepsilon \quad \text{for } Y=+1$$
$$X_0 := s(U)+\varepsilon = \begin{bmatrix}\cos(U)\\ \sin(U)\end{bmatrix}+\varepsilon \qquad \text{for } Y=-1.$$

For our examples we selected $\sigma = 0.2$ and $\lambda = 0.6$.

To re-sample the label Y after providing recourse we calculate the conditional distribution of this model. Let, p be the density of ε. Then, the density of $X_0 \mid Y=1$ and $X_0 \mid Y=-1$, denoted by g_+ and g_- respectively, is given by

$$g_+(x_0) = \frac{1}{2\pi}\int_0^{2\pi} p(x_0 - \lambda s(u))\mathrm{d}u,$$
$$g_-(x_0) = \frac{1}{2\pi}\int_0^{2\pi} p(x_0 - s(u))\mathrm{d}u.$$

Table 3: Estimated risks on the Credit dataset. Lower risk is bold.

	Wachter		GS		CoGS	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
LR	0.17 ± 0.04	**0.05 ± 0.02**	0.17 ± 0.04	**0.05 ± 0.01**	0.17 ± 0.03	**0.05 ± 0.01**
GBT	**0.07 ± 0.00**	**0.07 ± 0.00**	**0.07 ± 0.00**	**0.07 ± 0.00**	**0.07 ± 0.00**	**0.07 ± 0.00**
DT	**0.27 ± 0.09**	**0.12 ± 0.18**	0.23 ± 0.13	**0.05 ± 0.01**	0.23 ± 0.11	**0.06 ± 0.01**
NB	**0.18 ± 0.27**	**0.06 ± 0.01**	**0.16 ± 0.17**	**0.06 ± 0.01**	**0.18 ± 0.37**	**0.06 ± 0.01**
QDA	**0.21 ± 0.39**	**0.07 ± 0.01**	**0.14 ± 0.14**	**0.06 ± 0.01**	**0.13 ± 0.15**	**0.07 ± 0.02**
NN 1	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.00**
NN 2	**0.07 ± 0.01**	**0.06 ± 0.01**	**0.07 ± 0.01**	**0.07 ± 0.01**	**0.07 ± 0.01**	**0.07 ± 0.01**
NN 3	**0.07 ± 0.01**	**0.07 ± 0.01**	**0.07 ± 0.00**	**0.07 ± 0.00**	**0.07 ± 0.01**	**0.07 ± 0.00**
NN 4	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.00**	**0.07 ± 0.00**
NN 5	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.01**	**0.06 ± 0.00**	**0.07 ± 0.00**

Table 4: Estimated risks on the HELOC dataset. Lower risk is bold.

	Wachter		GS		CoGS	
	R_P	R_Q	R_P	R_Q	R_P	R_Q
LR	**0.27 ± 0.03**	0.40 ± 0.04	**0.28 ± 0.02**	0.42 ± 0.02	**0.27 ± 0.04**	0.45 ± 0.02
GBT	**0.19 ± 0.02**	**0.21 ± 0.04**	**0.20 ± 0.02**	**0.25 ± 0.12**	**0.20 ± 0.01**	0.35 ± 0.05
DT	**0.19 ± 0.01**	**0.29 ± 0.31**	**0.20 ± 0.02**	**0.25 ± 0.13**	**0.20 ± 0.02**	0.35 ± 0.09
NB	**0.29 ± 0.02**	0.45 ± 0.03	**0.29 ± 0.02**	0.45 ± 0.07	**0.28 ± 0.03**	0.51 ± 0.05
QDA	**0.32 ± 0.03**	0.47 ± 0.03	**0.32 ± 0.02**	0.49 ± 0.03	**0.31 ± 0.03**	0.52 ± 0.03
NN 1	**0.27 ± 0.03**	0.46 ± 0.03	**0.28 ± 0.02**	0.46 ± 0.03	**0.28 ± 0.02**	0.49 ± 0.02
NN 2	**0.30 ± 0.12**	0.47 ± 0.02	**0.27 ± 0.03**	0.45 ± 0.04	**0.30 ± 0.12**	0.51 ± 0.07
NN 3	**0.28 ± 0.03**	0.46 ± 0.04	**0.27 ± 0.02**	0.46 ± 0.02	**0.27 ± 0.03**	0.50 ± 0.02
NN 4	**0.27 ± 0.02**	0.44 ± 0.05	**0.27 ± 0.04**	0.45 ± 0.03	**0.26 ± 0.02**	0.48 ± 0.04
NN 5	**0.27 ± 0.02**	0.44 ± 0.05	**0.27 ± 0.04**	0.45 ± 0.03	**0.26 ± 0.02**	0.48 ± 0.04

In our implementation this integral is approximated by a Riemann sum. The conditional distribution now follows,

$$
\begin{aligned}
&P(Y = 1 \mid X_0 = x_0) \\
&\quad = \frac{g_+(x_0)P(Y = 1)}{g_+(x_0)P(Y = 1) + g_-(x_0)P(Y = -1)} \\
&\quad = \frac{g_+(x_0)}{g_+(x_0) + g_-(x_0)}.
\end{aligned}
$$

Gaussians dataset The Gaussian data points are sampled from 2 Gaussians with different means $\mu, \nu \in \mathbb{R}^2$ and different covariances $\Sigma_+, \Sigma_- \in \mathbb{R}^{2\times 2}$,

$$
\begin{aligned}
X \mid Y = +1 &\sim \mathcal{N}(\mu, \Sigma_+) \\
X \mid Y = -1 &\sim \mathcal{N}(\nu, \Sigma_+).
\end{aligned}
$$

The densities of the 2 conditional distributions are given by

$$
\begin{aligned}
g_+(x_0) &= \frac{1}{2\pi|\Sigma_+|} e^{-\frac{1}{2}(x_0-\mu)^\top \Sigma_+^{-1}(x_0-\mu)} \\
g_0(x_0) &= \frac{1}{2\pi|\Sigma_-|} e^{-\frac{1}{2}(x_0-\nu)^\top \Sigma_-^{-1}(x_0-\nu)}.
\end{aligned}
$$

The conditional distribution is given by the odds between the densities of the 2 gaussians

$$
P(Y = 1|X_0 = x_0) = \frac{g_+(x_0)}{g_-(x_0) + g_+(x_0)}.
$$

E.1.1 Linear relation between p and R_Q

Here, we derive the linear relation seen in Figure 4. Denote by $R_Q(p)$ the risk after recourse dependent on p, then the following relation can be calculated

$$
R_Q(p) = R_P(f) + p(R_Q(1) - R_P(f)). \tag{23}
$$

We can apply Lemma 7 and use that $r(x_0) = p$ for all x_0 to get the following expressions:

$$
\begin{aligned}
R_Q(p) &= \mathbb{E}\left[r(X_0)p_-(\varphi(X_0))\right] \\
&\quad + \mathbb{E}\left[(1 - r(X_0))1\{f(X_0) \neq Y\}\right] \\
&= p\mathbb{E}\left[p_-(\varphi(X_0))\right] + (1-p)\mathbb{E}\left[1\{f(X_0) \neq Y\}\right] \\
&= p\mathbb{E}\left[p_-(\varphi(X_0))\right] + (1-p)R_P.
\end{aligned}
$$

This also gives us that $R_Q(1) = \mathbb{E}\left[p_-(\varphi(X_0))\right]$. Substituting and rewriting gives

$$
\begin{aligned}
R_Q(p) &= pR_Q(1) + (1-p)R_P \\
&= R_P + p(R_Q(1) - R_P).
\end{aligned}
$$

E.2 Real Data

Here, we describe how the experiments for the real data were performed.

Conditional distribution estimation As mentioned before the main challenge with real data is that we do not have access to $P(Y \mid X_0)$. To circumvent this, we estimate this function as well as possible, by reserving most of the data to train a calibrated classifier. $N_{\text{cond train}}$ are used to train this classifier and $N_{\text{cond calib}}$ are used to calibrate this classifier. The exact values of the data splits are given in Table 5. Furthermore, we perform a grid search over a large set of parameters using cross validation to find the best performing calibrated classifier. The parameters in the grid search are

- `learning_rate`: $\{0.05, 0.15\}$,
- `n_estimators`: $\{10, 20, 60\}$,
- `subsample`: $\{0.8, 0.9, 1\}$,
- `max_depth`: $\{1, 2, 3\}$.

As a base classifier we use the `Gradient BoostedClassifier` from `scikit-learn` and we use Platt scaling (Platt et al., 1999) to calibrate the probabilities.

Table 5: Details of the datasets used during the experiments

	Credit data	Adult data	HELOC data
$P(Y = +1)$	0.932	0.239	0.480
$P(Y = -1)$	0.068	0.761	0.520
$N_{\text{cond train}}$	40000	30000	5000
$N_{\text{cond calib}}$	10000	10000	2000
N_{train}	5000	5000	5000
N_{test}	1000	1000	1000

Classification and Recourse After a conditional distribution is estimated for each dataset, we train the same set of classifiers as for the synthetic data on N_{train} different data points. Then, counterfactuals are generated using the different methods and using the trained conditional estimated distribution a new class label is sampled for the position at the counterfactual point. The estimated risk is then calculated for the dataset before and after recourse is provided.

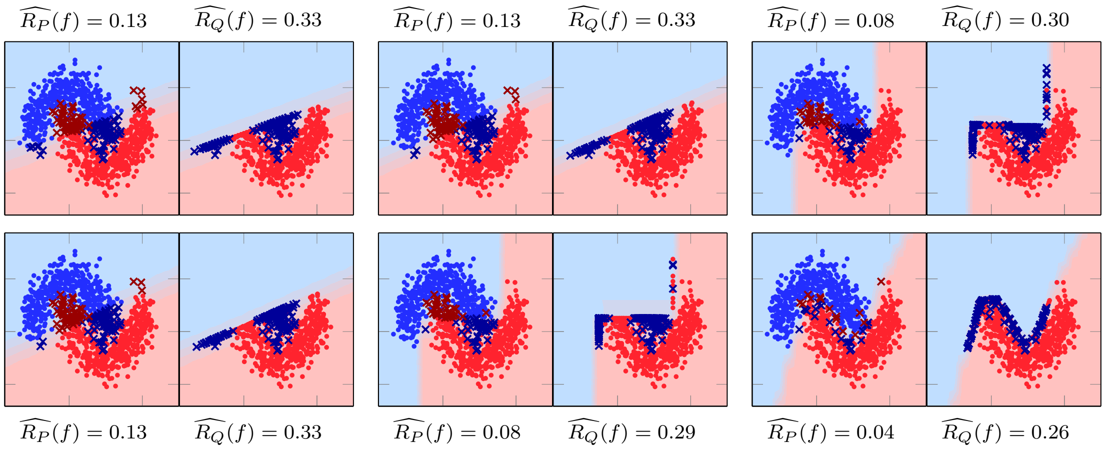

Figure 6: Examples of the effect of giving recourse with various classifiers on the Moons data set. From left to Right, Top to Bottom: LR, QDA, GBT, NB, DT, NN(8, 16).

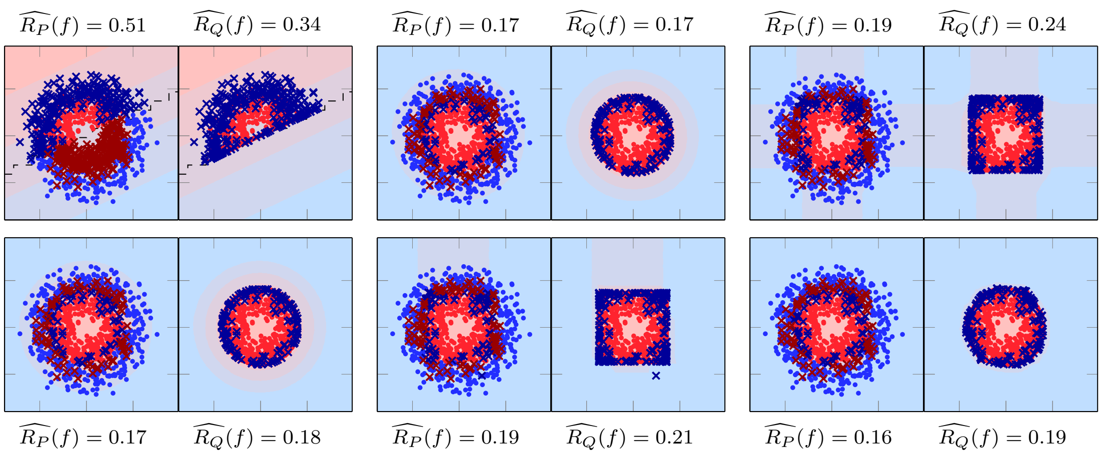

Figure 7: Examples of the effect of giving recourse with various classifiers on the Circles data set. From left to Right, Top to Bottom: LR, QDA, GBT, NB, DT, NN(8, 16).

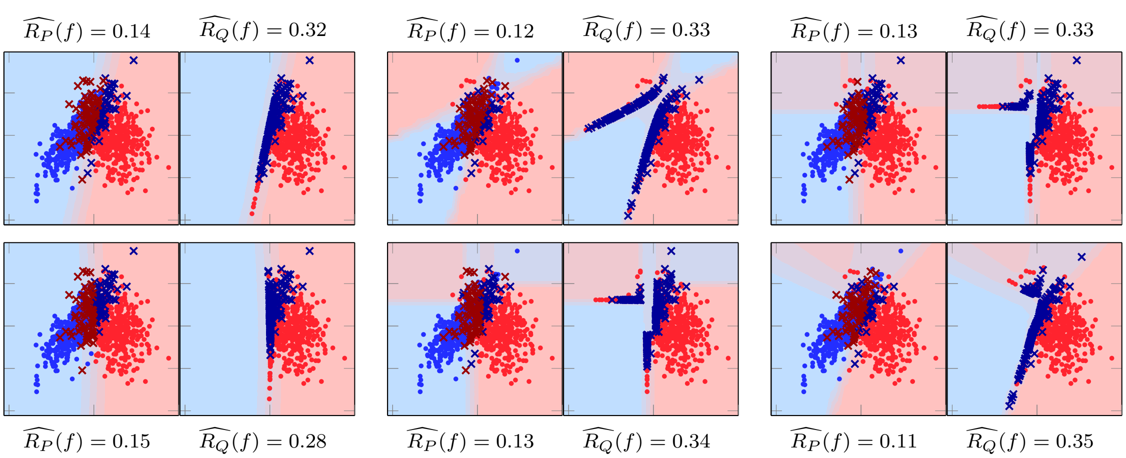

Figure 8: Examples of the effect of giving recourse with various classifiers on the Gaussians data set. From left to Right, Top to Bottom: LR, QDA, GBT, NB, DT, NN(8, 16).