Journal of Machine Learning Research 24 (2023) 1-37 Submitted 01/23; Revised 11/23; Published 11/23

Attribution-based Explanations that Provide Recourse Cannot be Robust

Hidde Fokkema H.J.FOKKEMA@UVA.NL
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands

Rianne de Heide R.DE.HEIDE@VU.NL
Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands

Tim van Erven TIM@TIMVANERVEN.NL
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands

Editor: Maxim Raginsky

Abstract

Different users of machine learning methods require different explanations, depending on their goals. To make machine learning accountable to society, one important goal is to get actionable options for *recourse*, which allow an affected user to change the decision $f(x)$ of a machine learning system by making limited changes to its input x. We formalize this by providing a general definition of recourse sensitivity, which needs to be instantiated with a utility function that describes which changes to the decisions are relevant to the user. This definition applies to local attribution methods, which attribute an importance weight to each input feature. It is often argued that such local attributions should be robust, in the sense that a small change in the input x that is being explained, should not cause a large change in the feature weights. However, we prove formally that it is in general impossible for any single attribution method to be both recourse sensitive and robust at the same time. It follows that there must always exist counterexamples to at least one of these properties. We provide such counterexamples for several popular attribution methods, including LIME, SHAP, Integrated Gradients and SmoothGrad. Our results also cover counterfactual explanations, which may be viewed as attributions that describe a perturbation of x. We further discuss possible ways to work around our impossibility result, for instance by allowing the output to consist of sets with multiple attributions, and we provide sufficient conditions for specific classes of continuous functions to be recourse sensitive. Finally, we strengthen our impossibility result for the restricted case where users are only able to change a single attribute of x, by providing an exact characterization of the functions f to which impossibility applies.

Keywords: explainability, interpretability, algorithmic recourse, theory

1. Introduction

As machine learning (ML) is changing science and society in many ways, its trustworthiness and interpretability are coming under increasing scrutiny (Varshney, 2022; Molnar, 2022). Since most ML systems are not inherently interpretable, there have been many proposals to generate explanations that communicate relevant aspects of their internal workings (Linardatos et al., 2020; Samek et al., 2019). Which particular aspects are relevant, depends on the target audience and its goals (Arrieta et al., 2020; Varshney, 2022). We consider here the case that the target audience consists of users with corresponding feature vectors $x \in \mathbb{R}^d$, who are affected by the decisions $f(x)$ of an ML system. It is assumed that each user has some limited ability to change (a subset of) their features in x, and the goal of the users is to use this ability to influence the resulting decision of f in a way that increases their utility by a sufficient amount. In order to achieve this goal, an explanation for a given input x should provide actionable options for *recourse*, i.e. changes to x that both provide sufficient utility *and* lie within the ability of the user to realize.

Attribution Methods and Counterfactual Explanations We consider explanations that take the form of local attributions $\varphi_f(x) \in \mathbb{R}^d$, which are vectors that assign a weight to each feature in x that indicates its importance. Many explanation methods produce such attributions. For instance, well-known methods like LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Integrated Gradients (Sundararajan et al., 2017) and SmoothGrad (Smilkov et al., 2017) are attribution methods. When applied to image classification, where the features are pixels, the attribution vector $\varphi_f(x)$ is called a *saliency map* and can be visualized as a picture that highlights the most important pixels. An approach closely related to attributions, which is often considered in the context of recourse, is to provide counterfactual explanations (Karimi et al., 2021; Verma et al., 2020; Keane et al., 2021). Methods of this type include those by Poyiadzi et al. (2020); Karimi et al. (2020); Wachter et al. (2017); Mothilal et al. (2020); Dandl et al. (2020); Dhurandhar et al. (2018). These methods generate an alternative (counterfactual) input x^{cf} that is both similar to x and provides sufficient utility. For example, if f is a classifier, then $f(x^{\text{cf}})$ might be a more desirable class for the user than $f(x)$. The differences between x^{cf} and x are then interpreted as the changes needed to flip the class, so that $\varphi_f(x) = x^{\text{cf}} - x$ can again be regarded as an attribution vector.

Robustness to Changes in the Inputs Developing precise design criteria for explainability methods has proven to be difficult (Jacovi and Goldberg, 2020; Zhou et al., 2021; Hooker et al., 2019). In the absence of these, a way forward is to consider desirable criteria that should be satisfied. One such criterion, which is commonly proposed in the context of recourse, is for explanations to be robust to changes in the inputs (Karimi et al., 2021; Alvarez-Melis and Jaakkola, 2018): if we reason that similar users should get similar options for recourse, then small changes in the input x should not cause large jumps in the explanation $\varphi_f(x)$. This can either be formalized to mean that φ_f should be continuous or, more restrictively, that it should be (locally) Lipschitz continuous. We will adopt the weaker of these two, because that strengthens our main theoretical results: a function that violates continuity automatically also violates Lipschitz continuity. Thus, we settle on the following definition of robustness:

Definition 1 *An attribution method $\varphi_f \colon \mathcal{X} \to \mathbb{R}^d$ is called* robust *if it is continuous.*

Robustness can be seen as a measure of coherence (Jacovi et al., 2023). It appears to be difficult to achieve, however, because a sequence of empirical counterexamples have been found in which several methods like LIME, SHAP and Integrated Gradients are not robust (Ghorbani et al., 2019; Slack et al., 2020; Dombrowski et al., 2019; Alvarez-Melis and Jaakkola, 2018).[1] On the other hand, it has been established that SmoothGrad and C-LIME (a continuous variant of LIME) *are* provably robust, because they produce attributions φ_f that are always Lipschitz continuous (Agarwal et al., 2021).

Main Contributions We take a more abstract look at why existing attribution methods may fail at being robust to changes in the inputs. Our explanation is that there is in fact a fundamental contradiction between robustness and providing recourse. This is established by our main result, Theorem 4 in Section 3, which shows that:

For any way of measuring utility, there exists a (continuous) machine learning model f for which no attribution method φ_f can be both recourse sensitive and continuous.

Our result captures many possible variations of how recourse may be defined via a permissive property we call *recourse sensitivity*. Recourse sensitivity is introduced in Section 2. In particular, it allows attributions $\varphi_f(x)$ to be scaled arbitrarily (which makes it easier for a suitable φ_f to exist), as long as the vector $\varphi_f(x)$ points in any direction that would allow the user to obtain sufficient utility. We also do not restrict to the case where users want to flip the class of a classifier f, but allow for a general utility function u_f that measures the user's utility. Finally, the contradiction between recourse sensitivity and continuity of φ_f does not require f to be some obscure function, but already occurs, for instance, for quadratic functions f (see Section 3.4). This implies that most model classed used in practice are expressive enough to exhibit the problem. In Section 3, we further illustrate our main impossibility result with experiments and analytical examples that show cases in which the well-known attribution methods SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) fail to be recourse sensitive. We also provide an analytical example in which counterfactual explanations fail to be continuous. We then reflect on our impossibility result in Section 4, and discuss possible ways around it. While the impossibility result implies that some functions f are problematic, it is still possible that joint recourse sensitivity and continuous attributions are possible under (necessarily restrictive) conditions on f, for instance if f is linear. We study this in Section 5, where we provide sufficient conditions on f that do generalize beyond the linear case. Finally, in Section 6, we strengthen the impossibility result from Theorem 4 and the sufficient conditions from Section 5 by providing an exact characterization of the set of functions f to which impossibility applies for two restricted special cases: first we characterize impossibility for dimension $d = 1$; then we extend this result to any $d \geq 1$ under the assumption that the user is only able to change a single feature.

1. Integrated Gradients is continuous by design, so in this case we can take the empirical results to mean that it is not Lipschitz continuous. In practice, the two notions are difficult to distinguish: if we only have sample access to a function, it is difficult to discern whether a sudden jump between two samples is caused by an actual discontinuity, or occurs because the function is very steep.

1.1 Related Work

The Taxonomy of Explanation Methods Interpretability of machine learning methods can be achieved by training inherently interpretable models f or by providing *post-hoc* explanations of a model f that has already been trained. Explanations can be *global*, explaining aspects of the full function f, or *local*, explaining the behavior of f around a given point x (Zhou et al., 2021; Molnar, 2022; Varshney, 2022; Das and Rad, 2020). Recourse fits into this taxonomy as a post-hoc, local type of explanations (Linardatos et al., 2020; Samek et al., 2019).

Distance Measures The survey by Karimi et al. (2021) provides a unified view on existing algorithms that provide recourse via counterfactual explanations. In the simplest case, such methods measure the distance between x and x^{cf} by the Euclidean distance, but more refined distance measures have also been proposed (Wachter et al., 2017; Karimi et al., 2020; Poyiadzi et al., 2020; Joshi et al., 2019; Arvanitidis et al., 2021). For simplicity, we restrict attention to the Euclidean distance in our results, but we expect that they can be generalized to many other distance measures as well.

Consequential Recommendations Karimi et al. (2021) further describe a generalization of counterfactual explanations, called *consequential recommendations*, in which users are not able to change individual features directly, but can only influence features indirectly via more abstract actions. The effect of actions on features is described by a causal model, and instead of the change in features the users are restricted by the cost of taking particular actions. As will be discussed in Section 2, our definition of recourse sensitivity is sufficiently general that it can also express consequential recommendations. Since the conditions of our main theorem are very mild, they will then also apply to many, but not all, causal models.

Other Notions of Robustness As mentioned already, robustness can mean multiple things. So far, we discussed (local) Lipschitz continuity and ordinary continuity. These notions both consider robustness with respect to the input variable x. One other notable interpretation of robustness is robustness of a counterfactual with respect to changes to the model f. For instance, a model may be periodically retrained (Ferrario and Loi, 2022) or it may be updated when someone wants to be removed from the training set (Pawelczyk et al., 2022). There have been multiple methods developed to generate counterfactuals that are still valid under these model shifts (Black et al., 2022; Upadhyay et al., 2021; Hamman et al., 2023). These types of robustness are orthogonal to the type of robustness we consider in this work.

Solidifying the Foundations of Explainability Explainability is an exciting new research area that is witnessing a flurry of new methods and ideas. But it is clear that no single explanation method can exist that is good for all purposes (Guidotti et al., 2018). As limitations of existing explanation methods are being discovered (Adebayo et al., 2018; Kindermans et al., 2019; Rudin, 2019), this has lead to a desire to solidify the foundations and practice of explainability research. For instance, there is a lively debate on how to measure desirable properties like faithfulness, fidelity, plausibility, etc. (Jacovi and Goldberg, 2020; Ge et al., 2021; Guidotti et al., 2018). In this context, it is important to know which desirable properties can coexist in principle, and our work contributes to this understanding by pointing out that robustness is incompatible with providing recourse.

A series of recent works find fault with post-hoc attribution methods, either by showing empirical examples in which they exhibit undesirable behavior or by establishing theoretical results that identify fundamental limitations. On the empirical side, Rudin (2019) gives multiple arguments why post-hoc methods should not be used for high-stake decisions, because they are often not faithful or provide too little detail; Laugel et al. (2019) show that post-hoc counterfactuals have a high risk of being far away from any ground truth data point; Slack et al. (2020) show that post-hoc methods can be used to hide biases in a model; and in an adversarial setting it has been shown that post-hoc explanations can easily be manipulated (Dombrowski et al., 2019; Bordt et al., 2022). On the theoretical side, subsequent to a pre-print of our work, Bilodeau et al. (2022) derived additional impossibility results, which apply to hypothesis testing between pairs of model behaviors from the output of attribution methods. They are able to obtain much more general conclusions than we do, because they restrict attention to the more narrow class of attribution methods that are complete and linear, whereas our impossibility results apply to any attribution method in general. A significant limitation of requiring completeness is that it excludes all counterfactual methods that are commonly used in algorithmic recourse.

Since many of the problems with attribution methods are related to challenges in describing exactly when and for what purpose explanation methods can be trusted, one way forward may lie in the calls for greater rigor by Lipton (2018) and Doshi-Velez and Kim (2017). Leavitt and Morcos (2020) even go as far as claiming that "interpretability research suffers from an over-reliance on intuition-based approaches that risk — and in some case have caused — illusory progress and misleading conclusions". These works plead for the development of theory that may lead to provably better interpretability methods, and we view our work as a contribution in that direction.

2. Recourse Sensitivity

In this section we formally introduce recourse sensitivity, and show that, on its own, it can always be satisfied, for instance by counterfactual explanations.

Setting We assume that x takes values in some domain $\mathcal{X} \subseteq \mathbb{R}^d$, and that the machine learning model f is an element of the set $\mathcal{F}$ of all functions from $\mathcal{X}$ to $\mathbb{R}$. An attribution method for a given function $f \in \mathcal{F}$ is a function $\varphi_f \colon \mathcal{X} \to \mathbb{R}^d$.

Utility Functions To define recourse sensitivity, we describe the user's preferences for a given model $f \in \mathcal{F}$ by a *utility function* $u_f \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the interpretation that $u_f(x, y)$ is the utility experienced by the user if they succeed in changing their original features x to new features y, which implies that the decision of machine learning model f changes from $f(x)$ to $f(y)$. We assume that the user is satisfied if they achieve utility $u_f(x, y) \geq \tau$, for some threshold $\tau \in \mathbb{R}$.[2] For instance, if f represents the score of one of the classes in a classification task, with the sign of f indicating whether the class is chosen by the classifier, then the user may have a preferred class they would like to be classified in. For example, the preferred class could be the class for which the score is positive. This objective, which is closely related to finding counterfactual explanations or adversarial

2. Mathematically, it is always possible to reduce to the case that $\tau = 0$ by subtracting τ from the utility function, but for simplicity we allow general τ.

examples (Karimi et al., 2021), can be described by

$$u_f(x, y) := f(y) \geq 0. \tag{1}$$

Alternatively, if f predicts a credit risk score for the user, they might care about increasing their score by some amount τ, which can be represented by

$$u_f(x, y) := f(y) - f(x) \geq \tau. \tag{2}$$

And, as a third example, if f outputs the probability of some event and the user would like to increase (or decrease) that probability by a certain percentage $p \times 100\%$, their goal could be expressed as

$$u_f(x, y) := \frac{f(y)}{f(x)} \geq 1 + p \tag{3}$$

(or $u_f(x, y) = \frac{f(x)}{f(y)} \geq 1/(1 - p)$).

In some of our formal results we will restrict attention to utility functions that depend on x and y only via the decisions $f(x)$ and $f(y)$ of f. This is a very natural restriction, which is satisfied by all examples (1), (2) and (3) given above.

Recourse Sensitivity Informally, we call an attribution method φ_f recourse sensitive if the user can always achieve sufficient utility by moving in the direction of the vector $\varphi_f(x)$. We aim for a very permissive definition, which covers all methods that can reasonably be said to provide recourse, so if there are multiple such directions at an input x, then we allow any direction; and if there is no such direction at x, then $\varphi_f(x)$ is allowed to be anything.

Formally, we assume the user is able to change their input x to an alternative input y over at most some distance $\delta \geq 0$. This can be used to express that a feature like income cannot double in a reasonable period of time. As discussed in the introduction, we restrict attention to Euclidean distance for simplicity. We further allow for additional constraints on the alternatives via a constraint set $C(x)$. Thus, the set of attainable points for a user with original input x is

$$A(x) = \{y \in \mathcal{X} \mid \|x - y\| \leq \delta, y \in C(x)\}.$$

The constraints $C(x)$ may express that the user is unable to change some features in x that they have no control over, like gender, age group or location (Poyiadzi et al., 2020; Mothilal et al., 2020), and we assume throughout that $x \in C(x)$, which means that the user always has the option of not changing their features. It could also be the case that the user can only change the features in a particular way (e.g. age can only increase) or that features can only be changed together, for instance as described by an underlying causal model as in consequential recommendations (see the introduction) (Karimi et al., 2021). In most of our results, the possible choices for $C(x)$ that we will focus on are:

(a) $C(x) = \mathcal{X}$: the unrestricted case;
(b) $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq k\}$ with $\|z\|_0$ denoting the number of coordinates in z that are non-zero: the sparse case in which changing each feature requires effort and it is assume that the user can change at most k features; and

(c) $C(x) = \{y \in \mathcal{X} \mid y = x + \alpha z, \alpha \geq 0, z \in D\}$ for some set of directions $D \subseteq \mathbb{R}^d$: the case that the user is only allowed to move in a restricted set of directions D.

The points around x that are both attainable by the user and provide sufficient utility to reach a given threshold τ then correspond to

$$T(x) = \{y \in A(x) \mid u_f(x, y) \geq \tau\}.$$

Our definition of recourse sensitivity now states that any attribution at x should point in the direction of some y that is in the set $T(x)$:

Definition 2 (Recourse Sensitivity) *Given a threshold $\tau \in \mathbb{R}$, constraint function C, and model $f \in \mathcal{F}$, an attribution method $\varphi_f \colon \mathcal{X} \to \mathbb{R}^d$ is called* recourse sensitive *if*

$$\varphi_f(x) = \alpha(y - x) \qquad \textit{for some } \alpha > 0 \textit{ and } y \in T(x),$$

for all $x \in \mathcal{X}$ for which $T(x)$ is non-empty.

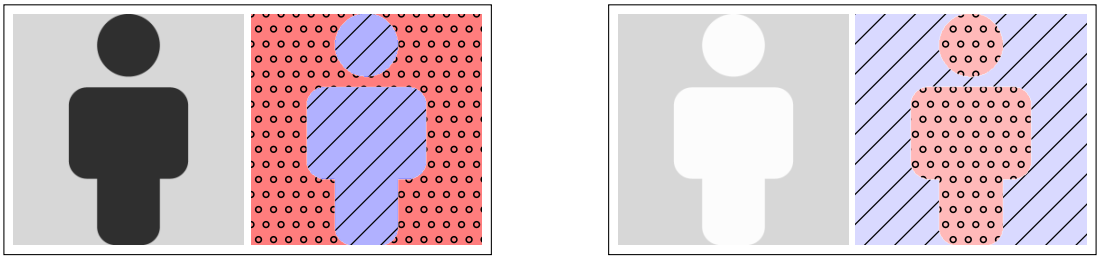
The case that $T(x)$ is empty corresponds to a user who has no options for recourse, so no explanation could possible help them. In this case we allow $\varphi_f(x)$ to be arbitrary.

Remark 3 (Satisfying Recourse Sensitivity) *It is clear that, in the absence of further requirements on φ_f, recourse sensitivity can trivially be satisfied by setting $\varphi_f(x) = y - x$ for some arbitrary $y \in T(x)$ whenever $T(x)$ is non-empty, and setting $\varphi_f(x) = 0$ otherwise. It is also satisfied by any counterfactual explanation x^{cf} that minimizes $\|x^{\text{cf}} - x\|$ subject to the constraint that $u_f(x, x^{\text{cf}}) \geq \tau$ and $x^{\text{cf}} \in C(x)$. To see this, note that, if $\|x^{\text{cf}} - x\| \leq \delta$, then $x^{\text{cf}} \in T(x)$, so $\varphi_f(x) = x^{\text{cf}} - x$ is a recourse sensitive choice. And if $\|x^{\text{cf}} - x\| > \delta$, then $T(x)$ is empty and $\varphi_f(x) = x^{\text{cf}} - x$ is also allowed.*

2.1 Profile Picture Example

We will now illustrate the concept of recourse sensitivity in a more concrete setting. Consider the following use case: a user has to upload an official profile picture of themselves x to a website to obtain a personalized card, which grants them access to some service.[3] The receiving party performs an automated check to verify if there is enough contrast between the brightness of the person and the background of the image, which is implemented by a function f that computes the squared difference between the average pixel values of the background and the average pixel values of the person. The picture is then accepted only if $f(x) \geq \lambda_{\text{thresh}}$ for some threshold parameter λ_{thresh}. Users whose picture is rejected want to submit a correct picture that is accepted, which would correspond to the utility function $u_f(x, y) = f(y) - \lambda_{\text{thresh}} \geq 0$. The amount by which the user is able to increase the contrast may be described by a suitable choice of δ, and optionally by describing constraints on how the user can manipulate the image via $C(x)$. Two examples with corresponding saliency maps are shown in Figure 1. Negative values indicate that a part of the picture should be darker, while positive values indicate that a part should be lighter. It can be seen that both saliency maps indicate the parts of the profile picture that have to be adjusted to increase the contrast, which makes them recourse sensitive: For the rejected picture, increasing the contrast is a direction that points towards sufficient

3. This example was inspired by the third author's frustrating attempts to obtain a new transport card for the Dutch railways. In reality he could not figure out why his picture was rejected.



(a) Accepted profile picture

(b) Rejected profile picture

Figure 1: Two examples of profile pictures (left in each subfigure) and their corresponding saliency maps (right in each subfigure). The red areas with circles indicate positive values, and the blue areas with lines correspond to negative values. In both cases changing the picture in the direction of the saliency map will result in a larger contrast between the person and background, which indeed moves it further into/towards the accepted class.

utility, and is therefore recourse sensitive. For the accepted picture, increasing the contrast would only strengthen the classification of the preferred class. Further details about Figure 1 are provided in Appendix A, and this example is continued in Section 3.3.

3. Impossibility of Recourse with Robustness in General

In this section we first present our main impossibility result, which shows that no attribution method can be both recourse sensitive and robust at the same time. Since no method can have both properties, it follows that there must exist counterexamples for every existing attribution method in which it violates at least one of the two. We illustrate our main result by explicitly constructing such counterexamples, both analytically for SmoothGrad, Integrated Gradients and Counterfactual Explanations, and empirically for LIME, SHAP, SmoothGrad and Integrated Gradients. At the end of the section we sketch the idea behind the proof of our impossibility theorem. The full proof is provided in Appendix B.

3.1 Main Impossibility Result

We restrict our attention to utility functions that only depend on x and y via the decisions of the machine learning model f, i.e. for which there exists a function $\widetilde{u}$ such that $u_f(x, y) = \widetilde{u}(f(x), f(y))$. This is a very natural restriction that covers all examples from Section 2.

Theorem 4 *Let $\delta > 0$ and $\tau \in \mathbb{R}$ be arbitrary, and let the constraint function $C(x)$ be any of the choices (a), (b) or (c) on p. 6. Furthermore, assume the utility function u_f is of the form $u_f(x, y) = \widetilde{u}(f(x), f(y))$, and that there exist $z_1, z_2 \in \mathbb{R}$ for which $\widetilde{u}(z_1, z_2) \geq \tau$ and $\widetilde{u}(z_1, z_1) < \tau$. Finally, assume that $\mathcal{X} \subseteq \mathbb{R}^d$ contains a line segment ℓ of length strictly larger than δ and such that $\ell \subseteq C(x)$ for all $x \in \ell$. Then there exists a continuous function $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and continuous.*

The required existence of z_1 and z_2 rules out the trivial case that the user can never achieve sufficient utility or will already receive sufficient utility without changing their input x. The existence of the

line segment ℓ is used to ensure there is enough room within the domain $\mathcal{X}$ and the constraints $C(x)$ to construct the counterexample. Although already very mild, the conditions of Theorem 4 can potentially be generalized, as discussed in Section 7. As an example of a setting that is covered by the theorem, we may for instance consider the classification setting in which the user wants to be classified in some preferred class, with a full domain and without any constraints. The conditions of the theorem are then all satisfied, for any $z_1 < 0 < z_2$, and the result simplifies to:

Corollary 5 (Unconstrained Classification) *Suppose $\mathcal{X} = \mathbb{R}^d$, $C(x) = \mathbb{R}^d$, $u_f(x, y) = f(y)$, $\tau \in \mathbb{R}$ and $\delta > 0$. Then there exists a continuous function $f\colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and continuous.*

3.2 Analytical Examples

We proceed to construct explicit analytical counterexamples f and utility functions u_f for the SmoothGrad and Integrated Gradients attribution methods. Since both attribution methods are always continuous, it follows from Theorem 4 that they cannot always be recourse sensitive, which is what the counterexamples demonstrate. Both examples are in dimension $d = 1$ and we will assume that there are no constraints, i.e. $\mathcal{X} = C(x) = \mathbb{R}$.

Example 1 (SmoothGrad) *Consider the function $f(x) = x^2$, the utility $u_f(x, y) = f(y) - f(x)$ and $\tau \in \mathbb{R}$, which expresses that the user wants to increase f by at least τ. The attribution given to each point by the SmoothGrad procedure will be*

$$\varphi_f^{SG}(x) = \mathbb{E}_{a\sim N(0,\sigma^2)}\left[f'(x+a)\right] = \mathbb{E}_{a\sim N(0,\sigma^2)}[2x + 2a] = 2x.$$

Here, $N(0, \sigma^2)$ denotes the normal distribution with mean 0 and some specified variance parameter $\sigma^2 > 0$. In almost all points x this attribution indeed points in a direction that increases $f(x)$. However, in the point $x = 0$ the attribution is 0, which does not provide meaningful recourse, because it does not tell the user that they can in fact increase $f(x)$ by changing x. Whenever $\delta \geq \sqrt{\tau}$ the user would be able to achieve a sufficient increase in utility by moving in any direction from $x = 0$, and this violates recourse sensitivity.

Example 2 (Integrated Gradients) *We examine $f(x) = e^{-x^2}$ and $u_f(x, y) = \frac{f(x)}{f(y)} \geq \tau$ for some threshold $\tau > 1$, which means the user wants to decrease $f(x)$ by a factor of at least τ. Also choose δ such that $\delta \geq \sqrt{\log \tau}$. If the user starts in $x = 0$, then this is possible by moving far away enough in both directions. Indeed if you would move to some $y \in \mathbb{R}$ from $x = 0$ with $\delta \geq |y| \geq \sqrt{\log(\tau)}$, it is possible to decrease f by the requested fraction τ within a δ-distance of $x = 0$, because then*

$$u_f(0, y) = \frac{1}{e^{-y^2}} = e^{y^2} \geq e^{\log(\tau)} = \tau. \tag{4}$$

We can explicitly calculate the attributions of the Integrated Gradients method, which depends on a baseline point x^0:

$$\varphi_f^{IG}(x) = (x - x^0) \int_0^1 f'\left(x^0 + \alpha(x - x^0)\right)\, \mathrm{d}\alpha = f(x) - f(x^0) = f(x) - 1,$$

where we have chosen $x^0 = 0$ as the baseline. Note that $\varphi_f^{IG}(x) < 0$ for all $x \neq 0$ and $\varphi_f^{IG}(0) = 0$. In this case, recourse is provided for $x < 0$, as moving towards the negative side is the fastest way to decrease the output of f. However, recourse is not provided for $x = 0$, for which $\varphi_f^{IG}(0) = 0$, but it should be non zero because of (4)*. There are more points which do not get recourse in this example. The points with $x > \delta$ can only decrease their output by moving to the right. To decrease the output by moving to the left would require moving a distance that is larger than δ. To see this, note that the function f is symmetric and to decrease $f(x)$ by moving to the left from some $x > \delta$ would at least require that $y < -x$, which is already a distance of 2δ away from x. As noted before, the attribution is always negative, so we find that φ_f^{IG} is also not recourse sensitive for the points $x > \delta$.*

Let us look at one last example. We will show that in some cases for binary classification, a counterfactual explanation cannot be continuous.

Example 3 (Counterfactual Explanations) *For this example set $\mathcal{X} = \mathbb{R}^2, u_f(x, y) = f(y), \tau = 0, \delta > 1$ and*

$$f(x) = \|x\| - 1.$$

In this example, the points within the unit circle are classified in the negative class and the points outside it as the positive class. The utility is such that if you are inside the circle, you want to move out of it. We can construct a simple counterfactual method by setting

$$x^{CF}(x) = \underset{\|y\| \geq 1}{\arg\min} \|x - y\|.$$

This optimization problem is well defined and has a unique solution for every point $x \neq 0$. There, the problem is still well-defined, but no unique solution exists. For the points with $\|x\| \geq 1$, the minimizer will be the point itself, and for $\|x\| < 1$, it will be $x/\|x\|$. Using these counterfactuals we can build a recourse sensitive attribution function by setting

$$\varphi_f(x) = \begin{cases} 0 & \text{if } \|x\| \geq 1, \\ \frac{x}{\|x\|} - x & \text{if } 0 < \|x\| < 1. \end{cases}$$

For the point $x = 0$ we now have many options, as the attribution is allowed to point to any point on the unit circle. However, no choice we make can produce an attribution that is continuous at $x = 0$, because the limit of $\varphi_f(x)$ is different when x approaches 0 along different lines.

3.3 Profile Picture Example (Continued)

In this section we complement our analytical examples from the previous section with empirical examples in which well-known attribution methods are recourse insensitive. We will demonstrate this for LIME, SHAP, SmoothGrad and Integrated Gradients, which can all be used to generate saliency maps. Our setup continues the profile picture example from Section 2.1. Since the gradient of f is linear in x in this example, both SmoothGrad and Integrated Gradients simplify, and both coincide with the Vanilla Gradients method $\varphi_f(x) = \nabla f(x)$ (Simonyan et al., 2014). We will refer to all three together as 'Gradient Methods'. LIME for images depends on a partition of the image into superpixels. We consider two variants. One variant is 'LIME manual', in which we provide the indices of the person as one super-pixel, and the indices of the background for a second super-pixel.

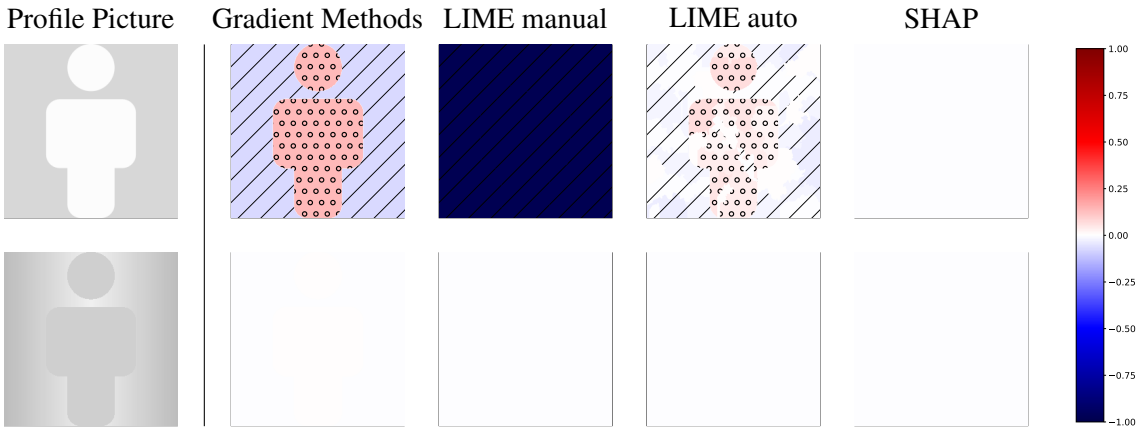


Figure 2: Saliency maps by different methods for two profile pictures that are rejected by the model.

The other variant is 'LIME auto', which uses the default segmentation algorithm of LIME to obtain the super-pixels. See Appendix A for further details about the experimental set-up.

Two example profile pictures with their corresponding saliency maps are shown in Figure 2. Both examples are rejected by the classifier. In both cases the user could change the decision by making the background darker (provided that δ is large enough). In the top row two of the methods, the Gradient Methods and LIME auto, do provide recourse. For LIME auto it is difficult to see, but the region of the person has a positive value and the surrounding region a negative value. Changing the picture accordingly would result in a darker background compared to the person, and therefore satisfies the requirement of recourse sensitivity. In the bottom example, the picture is also rejected. The attributions all show a flat saliency map, for which moving in the direction $\varphi_f(x)$ has no effect on the value of f, so none of the attribution methods is recourse sensitive. There is also significant disagreement between the saliency maps: LIME manual assigns a large positive value to all pixels, SHAP a very small almost unnoticeable negative value, and the remaining methods attribute 0 to all pixels.

3.4 Proof Idea for Theorem 4

Having illustrated the implications of Theorem 4 in the previous sections, we now explain the idea behind its proof. We consider again the setting of Example 1, with $f(x) = x^2$, $u_f(x, y) = f(y) - f(x)$ and $\delta \geq \sqrt{\tau} > 0$. Let L denote the interval of points x where recourse is possible by moving to the left, and let R denote the interval of points where recourse is possible by moving to the right. As illustrated in Figure 3, we then find that neither interval fully contains the other, but they overlap on $L \cap R = \left[\frac{\tau-\delta^2}{2\delta}, \frac{\delta^2-\tau}{2\delta}\right]$. Since the attribution has to be negative to the left of this overlap and positive to the right of the overlap, it has to go through zero somewhere inside the overlap (by continuity and the intermediate value theorem), but this is not allowed because the attribution is not allowed to be zero anywhere where recourse is possible. Therefore, no attribution can be continuous and recourse sensitive simultaneously. In the actual proof of Theorem 4, we follow a similar argument, but for a different function f, whose existence is guaranteed by the assumptions of

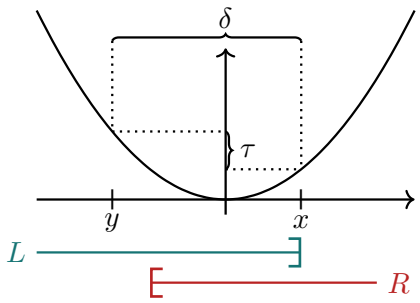


Figure 3: L and R sets for $f(x) = x^2$.

the Theorem. To generalize from one dimension to higher dimensions, we then embed this function along the line segment ℓ.

4. Discussion

In this section we pause to reflect on the implications of our impossibility result, and suggest possible ways to circumvent it.

How Common is the Problem? As is clear from the proof sketch in Section 3.4, recourse sensitivity and robustness get into conflict already for simple quadratic functions in 1D. And things are even worse: the same problem arises for any other function with partially overlapping L and R sets, so that on one part of the line recourse is provided by pointing to the left and on another part of the line recourse requires pointing to the right. We should therefore be prepared to run into problematic instances for most non-trivial learning models used in practice: e.g. linear models in which we add a quadratic feature, decision trees, neural networks, etc. In Sections 5 and 6 we provide more formal insights into the class of problematic functions f by providing conditions that are sufficient and/or necessary to combine recourse sensitivity and robustness.

Should We Prioritize Recourse Sensitivity or Robustness? In the context of algorithmic recourse, providing recourse is the primary goal and robustness is a secondary consideration. Faced with a choice between the two, it is therefore clear that we should prioritize recourse. However, it may be possible to (partially) salvage robustness in special cases, which seems important because it would be undesirable to have explanations that jump around unnecessarily. We proceed to explore this for counterfactual explanations.

4.1 Workarounds for Counterfactual Explanations

As pointed out in Remark 3, counterfactual explanations are always recourse sensitive. In Section 5 it will be shown that, under weak conditions, their robustness fails only at points x for which the counterfactual projection x^{cf} is not unique. This suggests two natural, but ultimately unsatisfactory ways to work around our impossibility result:

1. Robustness at Most Points If the points of non-uniqueness are sufficiently uncommon, then we may simply ignore them and accept a lack of robustness in such exceptional cases. It is indeed

tempting to believe that non-uniqueness might be rare enough. For instance, in the binary classification setting from (1), the set of points with a non-unique projection has measure 0 (see discussion below Theorem 6), which means that it is in a sense small compared to the ambient space $\mathcal{X}$. But, unfortunately, this is not sufficient to conclude that the set of affected users would also be small: around any point of discontinuity of φ_f there exists a whole neighborhood of users x who can find a nearby alternative x' that would result in a very different explanation. In extreme cases it is even possible that every user x in $\mathcal{X}$ would be close to a point of discontinuity of φ_f. We would therefore need to have a stronger restriction on the set of discontinuities before we can dismiss them as sufficiently uncommon.

2. Restrict to Very Simple Models A radical way to avoid discontinuities altogether is to restrict attention to very simple models f and constraint sets $C(x)$ for which the projection x^{cf} is always unique. For instance, if the user wants to be classified to a preferred class and there are no constraints, then this would be satisfied by functions f that are linear in the original features. Although effective, this approach is so restrictive that it seems unworkable, because non-uniqueness will quickly reappear, for instance if we add a quadratic feature as discussed at the start of the section.

4.2 Work-arounds by Changing the Explainability Task

More appealing options to avoid impossibility become available if we allow ourselves to change the explanation task. Some preliminary thoughts in this direction are as follows:

1. Linearizing with Abstract Features It may sometimes be an option to provide explanations not in terms of the original features x but in terms of transformed (typically more abstract) features $z = g(x)$ for some mapping g. If f is linear in z, then this would allow using a simple model after all. Appendix E.2 provides a detailed example to illustrate how this might work out.

2. Set-valued Explanations As discussed in Section 3.4, attributions cannot be continuous and recourse sensitive, because they may have to communicate different options in adjacent regions and there is no continuous way to transition between the options. A way around this may be to communicate not one, but many or all possible directions that provide recourse as a set $S_f(x)$. This is analogous to the definition of the subdifferential of a convex function, which represents the set of all possible tangents. Set-theoretic notions of continuity such as hemi-continuity or continuity with respect to the Hausdorff metric (Aubin and Frankowska, 2009) could then be used to rephrase robustness in terms of continuity of $S_f(x)$ instead of continuity of $\varphi_f(x)$, making robustness easier to satisfy.

5. Sufficient Conditions for Recourse with Robustness

The impossibility result in Theorem 4 implies that there exist continuous functions f for which no attribution method can both provide recourse and be robust. But this may still be possible if we restrict attention to specific functions f that are somehow nice enough. For instance, as mentioned in the introduction, linear classifiers do allow robust and recourse sensitive attribution functions when the goal is to move to a preferred class in binary classification. In this section, we will first formalize a generalization of this result to a slightly larger class of functions for the binary classification setting.

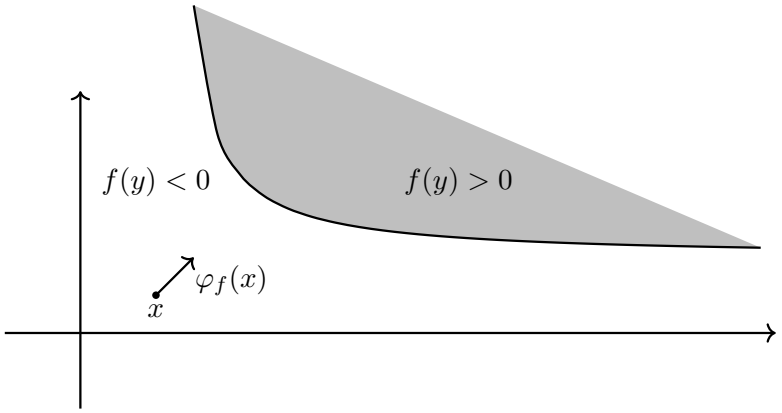


Figure 4: Assumptions of Theorem 6 illustrated.

Then we will extend the result to more general utilities. The proofs for this section can be found in Appendix C.

The following result shows that counterfactual explanations are both recourse sensitive and robust for binary classification with a preferred class if the counterfactual projections x^{cf} are always uniquely defined:

Theorem 6 *Consider the binary classification setting without constraints with $u_f(x, y) = f(y), \tau = 0$ and $C(x) = \mathcal{X}$, let $\delta > 0$ be arbitrary and take $f \colon \mathcal{X} \to \mathbb{R}$ to be any continuous function. If the set $U = \{y \in \mathcal{X} \mid f(y) \geq 0\}$ is convex, then the attribution method*

$$\varphi_f(x) \coloneqq \operatorname*{arg\,min}_{y \in U} \|y - x\| - x = P_U(x) - x$$

is well defined, and it is both recourse sensitive and continuous.

Theorem 6 covers linear functions f, but applies more broadly. For instance, any concave and continuous function f will satisfy its requirements. However, these requirements are still very restrictive, as only very simple classifiers will be concave.

To understand the conditions of Theorem 6, we note that continuity of f is used to ensure that U is closed, so that projections always exist. In this case the convexity condition on U is equivalent to the assumption that projections onto U are always unique, provided that $\mathcal{X}$ is convex. Thus the conditions of the theorem may also be understood as requirements to ensure unique projections. In light of the discussion in Section 4.1, we further remark that, without the convexity assumption on U, the set of points x with a non-unique projection onto U still has measure 0, as is shown by Erdös (1945) for general closed sets and $\mathcal{X} = \mathbb{R}^d$.

To extend Theorem 6 to more general utility functions and arbitrary constraints $C(x)$, we need to look at sets of the form

$$U(x) \coloneqq \{y \in \mathcal{X} \mid u_f(x, y) \geq \tau\} \cap C(x)$$

and consider counterfactual explanations that project onto these sets. Now the set that is projected on changes with x, so we will need more complicated assumptions to ensure that the projections still

change continuously with respect to x. Intuitively, this will be the case if the sets $U(x)$ themselves vary continuously with x. It turns out that the right type of set-valued continuity is *hemi-continuity*, which ensures that the sets $U(x)$ cannot explode, implode or shift suddenly, when varying x. The definition and relevant properties of hemi-continuity are reviewed in Appendix C. Compared to Theorem 6 we are further able to weaken the convexity assumption on $U(x)$ to the requirement that a unique projection exists for x instead of for all points. This leads to the following result, also proved in Appendix C:

Theorem 7 *Let $\delta > 0, \tau \in \mathbb{R}$, $f\colon \mathcal{X} \to \mathbb{R}$ be a function, $C(x)$ constraint sets and $u_f(x, y)$ a utility function with the following properties:*

1. *The set-valued map $U(x)$ is hemi-continuous; and*
2. *$U(x)$ is a closed set for every $x \in \mathcal{X}$.*

Then there exists at least one attribution method

$$\varphi_f(x) \in \left(\operatorname*{arg\,min}_{y \in U(x)} \|y - x\|\right) - x,$$

and any such method is recourse sensitive. Moreover, φ_f will be continuous on the restriction of $\mathcal{X}$ to points x for which $P_{U(x)}(x) := \arg\min_{y \in U(x)} \|y - x\|$ is unique.

To see that Theorem 6 follows from Theorem 7, we note that, if $U(x) = U$ is the same for all x, then $U(x)$ is always hemi-continuous. In addition, U is the pre-image of a closed set under f, and will therefore be closed if f is continuous. Finally, uniqueness of all projections (and therefore continuity) is a consequence of convexity of U.

Theorem 7 also covers other cases of interest. For instance, it is sufficient to require the following:

1. The function $u_f(x, y)$ is continuous;
2. For each $x \in \mathcal{X}$, the function $y \mapsto u_f(x, y)$ is concave;
3. The domain $\mathcal{X}$ is compact.

These conditions imply that all $U(x)$ are convex and compact (i.e. closed and bounded), and consequently all projections are unique. It can also be checked that they ensure that U is hemi-continuous. All requirements of Theorem 7 are therefore satisfied and a continuous and recourse sensitive attribution function exists.

Another example, which is covered by Theorem 7, but not by the previous sufficient conditions, is the following: suppose the user wants to at least double the outcome of the model, which can be expressed by taking $u_f(x, y) = \frac{f(y)}{f(x)}$ and $\tau = 2$. Then the following model f satisfies the conditions:

Corollary 8 *Let $\delta > 0$, $\tau = 2$, $\mathcal{X} = \mathbb{R}^d \setminus \{0\}$, $C(x) = \mathcal{X}$, $f(x) = e^{b\|x\|}$ for some $b > 0$ and $u_f(x, y) = \frac{f(y)}{f(x)}$. Then φ_f as defined in Theorem 7 is uniquely defined, recourse sensitive and continuous.*

To prove Corollary 8, it can be shown that $U(x) = \{y \in \mathcal{X} \mid \|y\| \geq \|x\| + \frac{\ln(2)}{b}\}$ is hemi-continuous. The sets $U(x)$ are also closed, and the projections onto $U(x)$ are unique for every x, because we

excluded 0 from our domain. Thus, all the requirements of Theorem 7 hold and a continuous and recourse sensitive attribution function exists.

6. Single Feature Recourse Sensitivity: Exact Characterization

In this section, we provide an in-depth study of the case that users are only able to change a single feature, which corresponds to the sparse constraint (b) from Section 2 with $k = 1$. In this case, we are able to provide an exact characterization of the functions f for which a continuous, recourse sensitive attribution function can exist. We first restrict ourselves to the one-dimensional case $d = 1$, and then generalize the result to the multi-dimensional case $d \geq 1$. All proofs of the results in this section can be found in Appendix B. To formulate our results, we will need the concept of separated sets, which corresponds to the intuition that the sets are disjoint and have at least one point in between them everywhere:

Definition 9 *Two sets $A, B \subseteq \mathcal{X}$ are called* separated *if $\mathrm{cl}(A) \cap B = \varnothing$ and $A \cap \mathrm{cl}(B) = \varnothing$.*

An equivalent definition is that there exist open sets $U, V \subseteq \mathcal{X}$ such that $A \subseteq U, B \subseteq V$ and $U \cap V = \varnothing$.

6.1 One Dimension

In this subsection we assume that $\mathcal{X} \subseteq \mathbb{R}$ is one-dimensional. Define the following three sets

$$\begin{aligned} L &= \{x \in \mathcal{X} \mid \text{ there exists some } y \in [x - \delta, x) \cap C(x) \text{ with } u_f(x, y) \geq \tau\}, \\ R &= \{x \in \mathcal{X} \mid \text{ there exists some } y \in (x, x + \delta] \cap C(x) \text{ with } u_f(x, y) \geq \tau\}, \\ O &= \{x \in \mathcal{X} \mid x \in C(x) \text{ and } u_f(x, x) \geq \tau\}. \end{aligned}$$

Similarly to Section 3.4, these sets are the feasible points for which recourse is obtainable by moving to the left, to the right or by doing nothing, respectively. The following result now tells us that a continuous recourse sensitive attribution function can exist if and only if we can decompose L, R and O in a particular way:

Theorem 10 *Let $\delta > 0, \tau \in \mathbb{R}, f \colon \mathcal{X} \to \mathbb{R}$ and $C(x)$ be arbitrary, then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exist $\widetilde{L} \subseteq L, \widetilde{R} \subseteq R$ and $\widetilde{O} \subseteq O$ such that*

1. *$\widetilde{L} \cup \widetilde{R} \cup \widetilde{O} = L \cup R \cup O$;*
2. *$\widetilde{L}$ and $\widetilde{R}$ are separated;*
3. *$\mathrm{cl}(\widetilde{O}) \cap \widetilde{L} = \varnothing$ and $\mathrm{cl}(\widetilde{O}) \cap \widetilde{R} = \varnothing$.*

Sufficient $\widetilde{L}$ and $\widetilde{R}$ sets Theorem 10 refers to the existence of any $\widetilde{L}, \widetilde{R}$ and $\widetilde{O}$ that satisfy its conditions, but we will proceed to show that it sufficient to check separatedness only for a restricted number of choices for $\widetilde{L}, \widetilde{R}$ and $\widetilde{O}$, which may even reduce to a single case. For simplicity, we will assume that $O = \varnothing$, so Condition 3 is automatically satisfied, but the result can be generalized to general O as well. To describe the choices for $\widetilde{L}$ and $\widetilde{R}$, it will be helpful to decompose L and R into the sets $\mathcal{L} = \{L_i \mid i \in \mathcal{I}\}$ and $\mathcal{R} = \{R_j \mid j \in \mathcal{J}\}$ of (maximal) intervals they contain, where these intervals may be open or closed on either side. Thus $L = \bigcup_{i \in \mathcal{I}} L_i$ and $R = \bigcup_{j \in \mathcal{J}} R_j$, and every

two distinct intervals $A, B \in \mathcal{L}$ (or $A, B \in \mathcal{R}$) are separated; otherwise they could be joined into a single larger interval $A \cup B$. We further note that in general the number of intervals in L or R may be uncountable, so the index sets $\mathcal{I}$ and $\mathcal{J}$ may be uncountable as well. Since splitting an interval in two will never result in two separated sets, we only need to make decisions about whole intervals. Moreover, most of these choices are forced, and we can define the remaining possibilities in terms of the following index sets:

$$\begin{aligned}\widetilde{I} &= \{i \in \mathcal{I} \mid \neg\exists j \in \mathcal{J} \text{ such that } L_i \subseteq R_j\},\\ \widetilde{J} &= \{j \in \mathcal{J} \mid \neg\exists i \in \mathcal{I} \text{ such that } R_j \subseteq L_i\},\\ \widetilde{K} &= \{i \in \mathcal{I} \mid \exists j \in \mathcal{J} \text{ such that } L_i = R_j\}.\end{aligned}$$

Theorem 11 *Let $\delta > 0, \tau \in \mathbb{R}$, $f\colon \mathcal{X} \to \mathbb{R}$ and $C(x)$ be arbitrary, and let u_f be any utility function with $u_f(x,x) < \tau$ for all $x \in \mathcal{X}$. Then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exists a partition $\{\widetilde{K}_1, \widetilde{K}_2\}$ of $\widetilde{K}$ such that the sets*

$$\widetilde{L} = \Big(\bigcup_{i\in\widetilde{I}} L_i\Big) \cup \Big(\bigcup_{i\in\widetilde{K}_1} L_i\Big) \quad \text{and} \quad \widetilde{R} = \Big(\bigcup_{j\in\widetilde{J}} R_j\Big) \cup \Big(\bigcup_{i\in\widetilde{K}_2} L_i\Big) \tag{5}$$

are separated.

Thus, the only choice in selecting $\widetilde{L}$ and $\widetilde{R}$ is how to divide the intervals indexed by $\widetilde{K}$, which appear both in L and R. In particular, if $\widetilde{K}$ is empty, then so are $\widetilde{K}_1$ and $\widetilde{K}_2$, and we only need to check separatedness for a single choice of $\widetilde{L}$ and $\widetilde{R}$.

6.2 Higher Dimensions

It is possible to extend Theorem 10 to higher dimensions, i.e. $\mathcal{X} \subseteq \mathbb{R}^d$, whenever the user has control over only one feature at the same time. The constraint set in this case becomes $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq 1\}$. Analogously with the one dimensional case, we first define sets on which the attribution is allowed to be positive or negative in the i'th feature. These sets are

$$\begin{aligned}L^i &= \{x \in \mathcal{X} \mid \text{ there exists some } \mathcal{X} \ni y = x - \alpha e_i, \alpha \in (0,\delta], \text{ such that } u_f(x,y) \geq \tau\},\\ R^i &= \{x \in \mathcal{X} \mid \text{ there exists some } \mathcal{X} \ni y = x + \alpha e_i, \alpha \in (0,\delta], \text{ such that } u_f(x,y) \geq \tau\},\\ O &= \{x \in \mathcal{X} \mid u_f(x,x) \geq \tau\}.\end{aligned}$$

Where e_i denotes the i'th basis vector of the standard basis. If the sets L^i, R^i and O can be decomposed in a way similar to the decomposition in Theorem 10, then a continuous recourse sensitive attribution function exists in the higher dimensional case.

Theorem 12 *Let $\delta > 0, \tau \in \mathbb{R}$, $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq 1\}$ and $f\colon \mathcal{X} \to \mathbb{R}$ be arbitrary. Then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exist $\widetilde{L}^i \subseteq L^i$ and $\widetilde{R}^i \subseteq R^i$ for all $i = 1, \ldots, d$ and $\widetilde{O} \subseteq O$ such that*

1. *$\widetilde{O} \cup \bigcup_{i=1}^d (\widetilde{L}^i \cup \widetilde{R}^i) = O \cup \bigcup_{i=1}^d (L^i \cup R^i)$;*
2. *All sets in $\{\widetilde{L}^1, \widetilde{R}^1, \ldots, \widetilde{L}^d, \widetilde{R}^d\}$ are pairwise separated;*
3. *$\mathrm{cl}(\widetilde{O}) \cap \widetilde{L}^i = \varnothing$ and $\mathrm{cl}(\widetilde{O}) \cap \widetilde{R}^i = \varnothing$ for all $i = 1, \ldots, d$.*

7. Conclusion

We showed that there are machine learning models for which it is impossible for any attribution method to be both recourse sensitive and continuous (i.e. robust). This was illustrated by examples exhibiting the problem for specific attribution methods, and we gave an exact characterization of the set of problematic models for the case where the user is only able to make sparse changes that affect a single feature. It was further shown how, by making restrictive assumptions on f that satisfy certain sufficient conditions, it is possible to circumvent our impossibility result.

We view our work as a contribution to establishing solid foundational definitions for explainable machine learning. To obtain these in the context of providing recourse, it would be of particular interest to follow up on possible solutions to work around our impossibility result, for instance along the lines discussed in Section 4. Another direction for future work would be to extend the characterizations from Section 6 to the case where the user can change multiple features. This would pose significant new technical challenges, because, in contrast to the single-feature case, there are then an infinite number of directions that an attribution can point to. In addition, the very general definition of the utility function results in very unstructured spaces of possible directions. It may therefore be needed to specialize to particular utility functions to make progress. Finally, we remark that we defined recourse sensitivity using the Euclidean distance, but our proofs hardly use any of its special properties. It should therefore be possible to extend our result to other distances, such as (weighted) combinations of ℓ_p norms or weighted Manhattan distance (Karimi et al., 2021), or to a setting where the norm is replaced by a (possibly asymmetrical) cost mechanism or causal mechanism.

Acknowledgments

The authors would like to thank Joris Bierkens for suggesting to add Theorem 6 and Royi Jacobovic for pointing out Berge's Maximum Theorem, which is key to the proof of Theorem 7. De Heide and Van Erven were supported by the Netherlands Organization for Scientific Research (NWO) under grant numbers 019.202EN.004 and VI.Vidi.192.095, respectively.

Appendix A. Details of Experimental Set-up

All the code to reproduce the experiments and figures in this paper can be found in a GitHub repository[4]. All experiments were run locally on an Apple MacBook Pro M1 13", 2020 with 8GB of RAM.

A.1 Profile Picture Toy Dataset

A total of 53 gray scale figures were created from the **User Icon** picture, found on `www.iconarchive.com`. [5] Each figure consists of two components, the person and a background. The figures have varying contrasts between these two components. We labeled each figure by hand according to

4. github.com/HiddeFok/recourse-robust-explanations-impossible
5. The icon is provided for free for non-commercial use.

this contrast. A figure with high enough contrast is labeled as "Accepted", while low contrast results in a "Rejected" label. The labeling was done in such a way that a perfect classifier exists, which is based on the quadratic difference between the mean pixel value of the person and the background. In the following expressions, $x \in \mathbb{R}^N$ denotes the vectorized version of a picture of size $N = wh$, where w and h are the width and height of the picture. The classification function is given by

$$f(x) = \left(\frac{1}{|I_{\text{per}}|} \sum_{i \in I_{\text{per}}} x_i - \frac{1}{|J_{\text{back}}|} \sum_{j \in J_{\text{back}}} x_j \right)^2 .$$

Where, I_{per} denotes the indices of the pixels belonging to the person, and J_{back} contains the indices of the background. A figure is accepted if $f(x) \geq \lambda_{\text{thresh}}$ for some threshold parameter λ_{thresh}. By increasing the threshold from the minimum value of all quadratic differences to the maximum value, the parameter with the highest accuracy was chosen. This lead to the choice $\lambda_{\text{thres}} = 5961.34$, which achieved perfect accuracy across both classes.

Several attribution methods were applied to each figure with this f. The methods used were Vanilla Gradients (Simonyan et al., 2014), SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017), LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). In Figure 5 six example pictures with their attributions are displayed. The attribution methods based on gradients were calculated analytically. The attributions for the Vanilla Gradients, SmoothGrad and Integrated Gradients are given by

$$\begin{aligned}
\varphi_f^{\text{VG}}(x)_k &= \begin{cases} \frac{2}{|I_{\text{per}}|} \left(\frac{1}{|I_{\text{per}}|} \sum_{i \in I_{\text{per}}} x_i - \frac{1}{|J_{\text{back}}|} \sum_{j \in J_{\text{back}}} x_j \right) & \text{if } k \in I_{\text{per}}, \\ \frac{-2}{|J_{\text{back}}|} \left(\frac{1}{|I_{\text{per}}|} \sum_{i \in I_{\text{per}}} x_i - \frac{1}{|J_{\text{back}}|} \sum_{j \in J_{\text{back}}} x_j \right) & \text{if } k \in J_{\text{back}}, \end{cases} \\
\varphi_f^{\text{SG}}(x)_k &= \mathbb{E}_{a \sim N(x, \sigma^2 I_d)} \left[\nabla f(a)_k \right] = \varphi_f^{\text{VG}}(x)_k, \\
\varphi_f^{\text{IG}}(x)_k &= (x_k - x_k^0) \int_0^1 \nabla f \left(x^0 + t(x - x^0) \right)_k \, \mathrm{d}t = \varphi_f^{\text{VG}}(\frac{1}{2}(x + x^0))_k,
\end{aligned}$$

where x^0 is some baseline picture. We choose x^0 to be the picture with all pixel values set to 0.

For LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), the libraries provided by their respective authors were used[6]. The LIME package is provided under the BSD 2-Clause License and SHAP is provided under the MIT License. For LIME we used version 0.2.0.1 and for SHAP version 0.40.0. The default parameters were used, unless specified otherwise. For the LIME method, superpixels are needed to create the attribution. We used two different methods to find these superpixels. In the 'LIME manual' method we manually supplied two superpixels: the first superpixel consists of the person and the second superpixel is the background. In the 'LIME auto' method we used the default segmentation algorithm. Finally, for some of the picture manipulation we used the scikit-image (van der Walt et al., 2011) package, version 1.0, under the BSD 3-Clause License[7].

6. Library for LIME: `https://github.com/marcotcr/lime`, library for SHAP: `https://github.com/slundberg/shap`
7. Scikit-image package: `https://github.com/scikit-image/scikit-image`

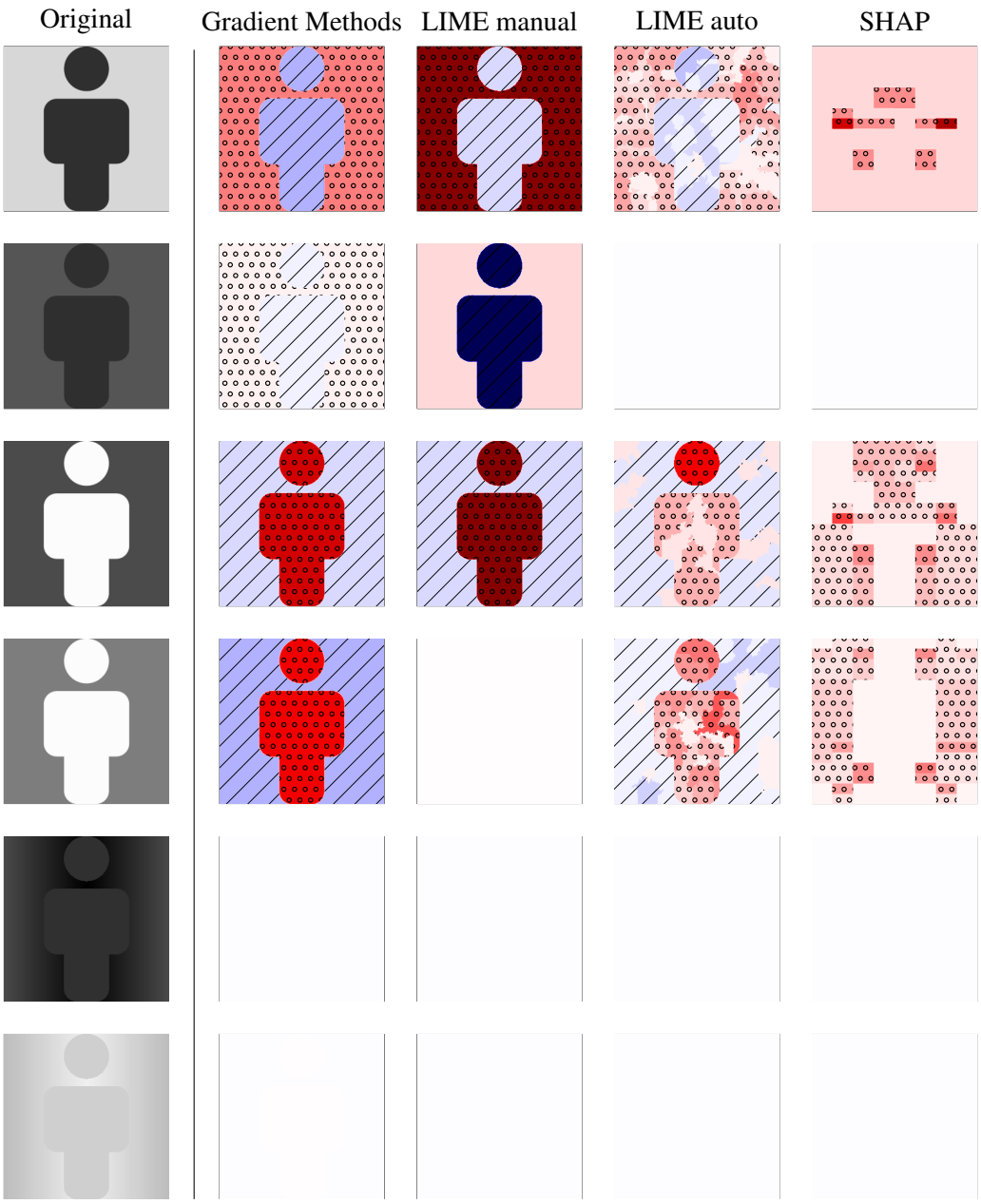


Figure 5: Additional examples of pictures and their attributions. From top to bottom the labels were: Accepted, Rejected, Accepted, Accepted, Rejected, Rejected.

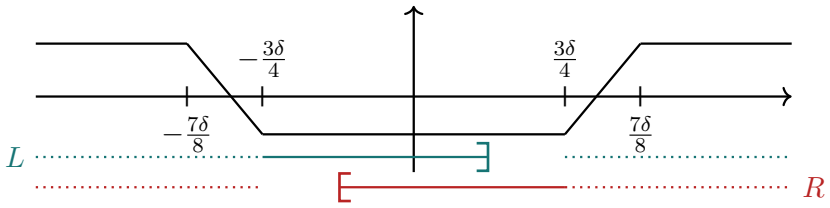


Figure 6: The constructed function $f(x)$. The dotted lines indicate that those values could be part of L or R, but do not have to be part of either set necessarily.

All proofs of the results in the main text can be found in the following sections. For clarity we will repeat the statements.

Appendix B. Proof of Theorem 4

Theorem 4 *Let $\delta > 0$ and $\tau \in \mathbb{R}$ be arbitrary, and let the constraint function $C(x)$ be any of the choices (a), (b) or (c) on p. 6. Furthermore, assume the utility function u_f is of the form $u_f(x, y) = \widetilde{u}(f(x), f(y))$, and that there exist $z_1, z_2 \in \mathbb{R}$ for which $\widetilde{u}(z_1, z_2) \geq \tau$ and $\widetilde{u}(z_1, z_1) < \tau$. Finally, assume that $\mathcal{X} \subseteq \mathbb{R}^d$ contains a line segment ℓ of length strictly larger than δ and such that $\ell \subseteq C(x)$ for all $x \in \ell$. Then there exists a continuous function $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and continuous.*

Proof We will split this proof into three parts. First, we consider the one-dimensional case, $\mathcal{X} = \mathbb{R}$. Then, we will show how to deal with $\mathcal{X} = \mathbb{R}^d$. Finally, we will discuss the result in its most general form, meaning $\mathcal{X} \subseteq \mathbb{R}^d$. Consider the case when $\mathcal{X} = \mathbb{R}$ and $C(x) = \mathcal{X}$. By assumption, there are two points $z_1, z_2 \in \mathbb{R}$ such that $\widetilde{u}(z_1, z_2) \geq \tau$. We can construct a continuous function f explicitly (see Figure 6):

$$f(x) = \begin{cases} z_1 & |x| < \frac{3}{4}\delta, \\ \frac{8(z_2 - z_1)}{\delta}|x| + (7z_1 - 6z_2) & \frac{3\delta}{4} \leq |x| \leq \frac{7\delta}{8}, \\ z_2 & |x| > \frac{7\delta}{8}. \end{cases}$$

We will apply Theorem 10 to show that no attribution method φ_f can be both recourse sensitive and continuous on this function f. To this end, we first note that, for $x \in [-\frac{3\delta}{4}, \frac{\delta}{8}]$ and $y = x - \delta$ we have that $u_f(x, y) = \widetilde{u}(f(x), f(y)) = \widetilde{u}(z_1, z_2) \geq \tau$. Which means that the attribution φ is allowed to point to the left on $[-\frac{3\delta}{4}, \frac{\delta}{8}]$. By a similar argument, we find that φ_f is allowed to point to the right on $[-\frac{\delta}{8}, \frac{3\delta}{4}]$. Furthermore, we also see that φ_f is not allowed to point towards the left on $[\frac{\delta}{4}, \frac{3\delta}{4}]$, because f does not change on $[x - \delta, x]$ if $x \in [\frac{\delta}{4}, \frac{3\delta}{4}]$. Analogously, it can be shown that φ_f cannot point towards the right on $[-\frac{3\delta}{4}, -\frac{\delta}{4}]$. With regard to Theorem 10, this means that $[\frac{\delta}{4}, \frac{3\delta}{4}] \subseteq \widetilde{R}$ and $[-\frac{3\delta}{4}, -\frac{\delta}{4}] \subseteq \widetilde{L}$ for any decomposition $\widetilde{L}, \widetilde{R}$ with $\widetilde{L} \cup \widetilde{R} = L \cup R$. Now, it is not possible to separate $[-\frac{3\delta}{4}, -\frac{\delta}{4}]$ from the interval $[-\frac{3\delta}{4}, \frac{\delta}{8}]$, which means that we would need $[-\frac{3\delta}{4}, \frac{\delta}{8}] \subseteq \widetilde{L}$. Similarly, we would need $[-\frac{\delta}{8}, \frac{3\delta}{4}] \subseteq \widetilde{R}$. It follows that $\widetilde{L}$ and $\widetilde{R}$ are not disjoint and in particular can never be separated. We conclude that for this continuous f no continuous recourse sensitive φ_f can exist. Note that this argument implies that no φ_f could exist on the interval $[-\delta, \delta]$. So, failing to provide recourse or be robust is a local issue.

We will now generalize the above argument to the setting where $\mathcal{X} = \mathbb{R}^d$ and the constraint is given by $C(x) = \{y \in \mathbb{R}^d \mid \|x - y\|_0 \leq k\}$ or $C(x) = \{y \in \mathbb{R}^d \mid y = x + \alpha z, \alpha \geq 0, z \in D\}$. Actually, these two versions of constraints can be dealt with simultaneously, by rotating the input space in the latter case in such a way that one of the vectors $z \in D$ lies alongside one axis. The argument for the one dimensional result can now be embedded in these cases. Namely, for $x = 0$, find the component that is allowed to change by the constraints. Call this the i'th component. We can define a similar function as the function above,

$$f(x) = \begin{cases} z_1 & |x_i| < \frac{3}{4}\delta, \\ \frac{8(z_2 - z_1)}{\delta}|x_i| + (7z_1 - 6z_2) & \frac{3\delta}{4} \leq |x_i| \leq \frac{7\delta}{8}, \\ z_2 & |x_i| > \frac{7\delta}{8}. \end{cases}$$

This function is again continuous and only changes in the i'th coordinate. Repeating the argument of the one-dimensional case, we see that an attribution is allowed to be negative in the i 'th component on $L^i = \mathbb{R}^{i-1} \times [-\frac{3\delta}{4}, \frac{\delta}{8}] \times \mathbb{R}^{d-i-1}$ and negative on $R^i = \mathbb{R}^{i-1} \times [-\frac{\delta}{8}, \frac{3\delta}{4}] \times \mathbb{R}^{d-i-1}$. Just as before, we also see that $\varphi_f^i(x)$ is necessarily negative on the set $\mathbb{R}^{i-1} \times [-\frac{3\delta}{4}, -\frac{\delta}{4}] \times \mathbb{R}^{d-i-1}$, but this set cannot be separated from L^i, which ensures that $\varphi_f(x)_i$ has to be negative on the whole of L^i. Alternatively, $\varphi_f(x)_i$ has to be positive on $\mathbb{R}^{i-1} \times [\frac{\delta}{4}, \frac{3\delta}{4}] \times \mathbb{R}^{d-i-1}$. Hence, also on the whole of R^i by the inability of separating R^i from this set. However, as L^i and R^i are not disjoint, this is a contradiction. Thus, no continuous attribution function can exist for f in higher dimensions.

Finally, we need to handle the multidimensional case that $\mathcal{X} \subseteq \mathbb{R}^d$. By assumption we have a line segment $\ell \subseteq \mathcal{X}$ with the property that $\ell \subseteq C(x)$ for all $x \in \ell$. We can apply a suitable transformation to the input space, so that we can fall back on our previous argument. This transformation is to first translate the line segment such that its middle point becomes the origin. Next, we apply a rotation such that the line segment lies along side the i'th axis. Call this translation and rotation M and ρ, respectively. The desired function now becomes

$$g(x) = f \circ \rho \circ M(x),$$

where f is the function of the precious case. The function g does not allow any continuous recourse sensitive attribution function, as f did not allow this on the line segment $[-\delta, \delta]$ in the i 't component and $g(\ell) = f \circ \rho \circ M(\ell) \supseteq f([-\delta, \delta])$. Now, if a continuous and recourse sensitive attribution function φ_g would exist for g, we could construct one for f as well. This is done by setting

$$\varphi_f(x) = \varphi_g \circ M^{-1} \circ \rho^{-1}(x).$$

The inverses exist and as translations and rotations do not change distances, this will be a continuous recourse sensitive attribution function for f, which was not possible. So, no continuous recourse sensitive attribution method can exist for g on $\mathcal{X} \subseteq \mathbb{R}^d$. ■

Appendix C. Proofs of Section 5

Theorem 6 *Consider the binary classification setting without constraints with $u_f(x, y) = f(y), \tau = 0$ and $C(x) = \mathcal{X}$, let $\delta > 0$ be arbitrary and take $f \colon \mathcal{X} \to \mathbb{R}$ to be any continuous function. If the*

set $U = \{y \in \mathcal{X} \mid f(y) \geq 0\}$ is convex, then the attribution method

$$\varphi_f(x) \coloneqq \arg\min_{y \in U} \|y - x\| - x = P_U(x) - x$$

is well defined, and it is both recourse sensitive and continuous.

Proof The function P_U is well defined by the fact that U is closed and convex. The set U is closed as it is the pre-image of a closed set under a continuous function and this ensures that the projection exists. Convexity of U guarantees uniqueness of the projection. Additionally, it is known that the projection is a continuous function if the projection exists and is unique. It follows that φ_f is also continuous. This leaves us to check that the map is recourse sensitive. If x is such that $f(x) \geq 0$, then $\varphi_f(x) = 0$, which is a valid attribution, since $u_f(x, x) = f(x) \geq 0$. So, assume $f(x) < 0$ and take $\alpha = 1$ in the definition of recourse sensitivity. Then either $\|P_U(x) - x\| > \delta$, in which case $T(x) = \emptyset$ and recourse sensitivity holds trivially, or $\|P_U(x) - x\| \leq \delta$ so that $P_U(x) \in T(x)$ because $P_U(x) \in U$ by definition, so φ_f is again recourse sensitive. We conclude that φ_f is both continuous and recourse sensitive. ■

As stated in the main text, we will need some additional tools from the field of multi-valued analysis to prove the general result. First, we will need a definition of continuity for set-valued expressions.

Definition 13 (Hemi-continuity) *For topological spaces $\mathcal{X}$ and $\mathcal{Y}$, a set-valued function $U : \mathcal{X} \to 2^{\mathcal{Y}}$ is called* upper hemi-continuous (UHC) *at $x_0 \in \mathcal{X}$ if, for any open $B \subseteq \mathcal{Y}$ with $U(x_0) \subseteq B$, there exists an open neighbourhood A of x_0 such that for all $x \in A$, $U(x)$ is a subset of B.*

A set-valued function $U : \mathcal{X} \to 2^{\mathcal{Y}}$ is called lower hemi-continuous (LHC) *at $x_0 \in \mathcal{X}$, if for any open set $B \subseteq \mathcal{Y}$ intersecting $U(x_0)$ there exists an open neighbourhood A of x_0 such that $U(x)$ intersects B for all $x \in A$*

If U is UHC and LHC at x_0, then U is called hemi-continuous at x_0. *If U is hemi-continuous at every $x_0 \in \mathcal{X}$, then U is called* hemi-continuous.

We will also need the following two Lemmas. The first relates UHC and LHC to normal continuity, when U is single-valued. The second tells us when the graph of U is a closed set.

Lemma 14 *If U is UHC or LHC at $x_0 \in \mathcal{X}$ and single-valued in some neighbourhood $\mathcal{N}$ around x_0 , then the function $f : \mathcal{N} \to \mathcal{Y}$ such that $U(x) = \{f(x)\}$ is a continuous function at x_0.*

Proof Take some sequence $\{x_n\}_{n=1}^{\infty}$ that converges to x_0. Recall that convergence is equivalent with the following. For any open neighbourhood B of x_0, there exists an $N \in \mathbb{N}$ such that $n \geq N$ implies that $x_n \in B$.

Start by assuming that U is UHC and single-valued. Take any open neighbourhood B of $f(x_0)$. By U being UHC, we can find an open neighbourhood A of x_0 such that $U(y) \subseteq B$ for all $y \in A$. Using the above characterisation of convergence, we can find an $N \in \mathbb{N}$ such that $x_n \in A$, whenever $n \geq N$. This also means that $\mathrm{B}\{f(x_n)\} = U(x_n) \subseteq B$. In particular, $f(x_n) \in B$. As B was arbitrary, it follows that $\lim_{n \to \infty} f(x_n) = f(x_0)$ and that f is continuous at x_0.

Next, we assume that U is LHC and single-valued. Again, take any any open set B such that $U(x_0) \cap B \neq \varnothing$. By the fact that $U(x_0)$ is single-valued, this actually means that $\{f(x_0)\} = U(x_0) \subseteq B$. By an analogous argument we again find that f is continuous at x_0. ■

Lemma 15 *If U is UHC and $U(x)$ is a closed set for all $x \in \mathcal{X}$, then the set*

$$\mathrm{Gr}(U) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid y \in U(x)\}$$

is closed.

Proof See Proposition 1.4.8 in Aubin and Frankowska (2009). ■

Theorem 16 (Berge's Maximum Theorem) *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, assume that:*

1. *The function $v : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function;*
2. *The set-valued function $U : \mathcal{X} \to 2^{\mathcal{Y}}$ is hemi-continuous, never empty, and assumes compact sets.*

Then, the parametrized optimization problem $v^*(x) \coloneqq \inf_{y \in U(x)} v(x, y)$ is continuous and the set-valued solution function $U^*(x) = \{y \in U(x) \mid v^*(x) = v(x, y)\}$ is UHC and compact-valued.

Proof See Chapter 6 in Berge (1997). ■

Now, Theorem 7 follows almost immediately from Theorem 16. The only issue is that the set $U(x)$ needs to be compact to apply Theorem 16. However, we do not want to impose this. Luckily, there exists a relaxation of Berge's Maximum Theorem, where we do not need compact-valued sets. This will require an additional property of the optimization problem, but this will be satisfied by the Euclidean norm.

Theorem 17 (Berge's Maximum Theorem for Non-Compact Image Sets) *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, assume that:*

1. *The function $v : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous and that for every compact $K \subseteq \mathcal{X}$ the set*

$$D_v(\lambda; K) = \{(x, y) \in K \times \mathcal{Y} \mid y \in U(x), v(x, y) \leq \lambda\}$$

   is compact for all $\lambda \in \mathbb{R}$;
2. *The set-valued $U : \mathcal{X} \to 2^{\mathcal{Y}}$ is LHC and never empty.*

Then, the parametrized optimization problem $v^*(x) = \inf_{y \in U(x)} v(x, y)$ is continuous and the solution set-valued function $U^*(x) = \{y \in U(x) \mid v^*(x) = v(x, y)\}$ is UHC and compact-valued.

At this point, we have all the tools required to prove Theorem 7. Let us repeat the statement.

Theorem 7 *Let $\delta > 0, \tau \in \mathbb{R}$, $f \colon \mathcal{X} \to \mathbb{R}$ be a function, $C(x)$ constraint sets and $u_f(x, y)$ a utility function with the following properties:*

1. *The set-valued map $U(x)$ is hemi-continuous; and*

2. *$U(x)$ is a closed set for every $x \in \mathcal{X}$.*

Then there exists at least one attribution method

$$\varphi_f(x) \in \left(\operatorname*{arg\,min}_{y \in U(x)} \|y - x\|\right) - x,$$

and any such method is recourse sensitive. Moreover, φ_f will be continuous on the restriction of $\mathcal{X}$ to points x for which $P_{U(x)}(x) := \operatorname{arg\,min}_{y \in U(x)} \|y - x\|$ is unique.

Proof We will split the proof into two parts. First, we will consider the case where the projection onto the sets $U(x)$ is actually unique for all $x \in \mathcal{X}$. Afterwards, we will discuss the case where the projection is not unique for every point.

We want to apply Theorem 17, where $v(x, y)$ is given by $v(x, y) = \|y - x\|$ and $\mathcal{X} = \mathcal{Y} \subseteq \mathbb{R}^d$. The set-valued U is given by all feasible points that achieve sufficient utility,

$$U(x) = \{y \in \mathcal{X} \mid u_f(x, y) \geq \tau\} \cap C(x).$$

The parametrized optimization problem will be given by $v^*(x) = \inf_{y \in U(x)} \|y - x\|$. By assumption, this infimum is attained, because $U(x)$ is closed, and it is unique. It rests to check that the sets $D_v(\lambda; K)$ are compact for all compact K and $\lambda \in \mathbb{R}$.

Let us decompose $D_v(\lambda; K)$ by setting

$$\begin{aligned} D_v(\lambda; K) &= \{(x, y) \in K \times \mathcal{Y} \mid y \in U(x), \|x - y\| \leq \lambda\} \\ &= \{(x, y) \in K \times \mathcal{Y} \mid y \in U(x)\} \cap \{(x, y) \in K \times \mathcal{X} \mid \|x - y\| \leq \lambda\}. \end{aligned}$$

The first of these sets can be further decomposed by intersecting $K \times \mathcal{Y}$ and $\mathrm{Gr}(U)$, for $\mathrm{Gr}(U)$ as defined in Lemma 15. The set $K \times \mathcal{Y}$ is closed, because it is the product of two closed sets, and the set $\mathrm{Gr}(U)$ is closed by Lemma 15, so their intersection must be closed as well. Similarly, it can be seen that the set $\{(x, y) \in K \times \mathcal{Y} \mid \|x - y\| \leq \lambda\}$ is closed, by writing it as the intersection between $K \times \mathcal{Y}$ and $\{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid \|x - y\| \leq \lambda\}$. The latter set is seen to be closed as it is the inverse image of closed set under a continuous function. Furthermore, the set $\{(x, y) \in K \times \mathcal{Y} \mid \|x - y\| \leq \lambda\}$ is a bounded set as it can be seen as the set K with a strip around it of size λ. It follows that $D_v(\lambda; K)$ closed and bounded, hence compact. As λ and K were arbitrary we see that $v(x, y) = \|x - y\|$ has the desired property.

As noted in the Theorem statement, the attribution function will be given by

$$\varphi_f(x) = \operatorname*{arg\,min}_{y \in U(x)} \|x - y\| - x = P_{U(x)}(x) - x.$$

We can now apply Theorem 17 and see that the solution $P_{U(x)}(x)$ is UHC and compact-valued. Furthermore, the projection does exist and is unique. Invoking Lemma 14 then tells us that $P_{U(x)}(x)$ is a continuous function. We see that $\varphi_f(x)$ is continuous and recourse sensitive by design.

Now, we will drop the assumption that every point $x \in \mathcal{X}$ has a unique projection onto $U(x)$. Consider the subset $X \subseteq \mathcal{X}$ for which each point does have projection onto $U(x)$. Now, we can repeat the proof shown above using Berge's Maximum theorem with the sets $\mathcal{X} = X$ and $\mathcal{Y} = \mathcal{X}$, as

the sets $U(x)$ will not be subsets of X in general. This will result in a continuous projection from X onto the sets $U(x)$ for all points in X and the function

$$\varphi_f(x) \coloneqq \underset{y \in U(x)}{\arg\min} \|x - y\| = P_{U(x)} - x$$

will be well defined, continuous and a recourse sensitive attribution function on the restricted set X. ■

Appendix D. Proofs of Section 6

Theorem 10 *Let $\delta > 0, \tau \in \mathbb{R}, f \colon \mathcal{X} \to \mathbb{R}$ and $C(x)$ be arbitrary, then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exist $\widetilde{L} \subseteq L, \widetilde{R} \subseteq R$ and $\widetilde{O} \subseteq O$ such that*

1. *$\widetilde{L} \cup \widetilde{R} \cup \widetilde{O} = L \cup R \cup O$;*
2. *$\widetilde{L}$ and $\widetilde{R}$ are separated;*
3. *$\mathrm{cl}(\widetilde{O}) \cap \widetilde{L} = \varnothing$ and $\mathrm{cl}(\widetilde{O}) \cap \widetilde{R} = \varnothing$.*

Proof *If*: Assume that $\widetilde{L} \subseteq L, \widetilde{R} \subseteq R$ and $\widetilde{O} \subseteq O$ exist with properties $(1) - (3)$. We will construct φ_f explicitly. To this end, we define the distance to a set A as

$$d(x, A) = \inf_{y \in A} |x - y|. \tag{6}$$

It is known that $d(x, A)$ is continuous for any set, see for example Chapter 2.5 in Mendelson (1990). By separatedness of $\widetilde{L}$ and $\widetilde{R}$ we can find open neighborhoods $U_1, V_1 \subseteq \mathbb{R}$ of $\widetilde{L}$ and $\widetilde{R}$ respectively, such that $U_1 \cap V_1 = \varnothing$. Furthermore, by property (3) we can also find other open neighbourhoods of $U_2, V_2 \subseteq \mathbb{R}$ of $\widetilde{L}$ and $\widetilde{R}$ such that $\widetilde{O} \cap U_2 = \varnothing$ and $\widetilde{O} \cap V_2 = \varnothing$. The sets $U = U_1 \cap U_2$ and $V = V_1 \cap V_2$ are still open neighbourhoods of $\widetilde{L}$ and $\widetilde{R}$ and they are disjoint from each other and $\widetilde{O}$. Now define

$$\varphi_f^-(x) = \frac{d(x, \mathbb{R} \setminus U)}{1 + d(x, \mathbb{R} \setminus U)},$$
$$\varphi_f^+(x) = \frac{d(x, \mathbb{R} \setminus V)}{1 + d(x, \mathbb{R} \setminus V)}.$$

Using these functions, we can construct φ by setting

$$\varphi_f(x) = \varphi_f^+(x) - \varphi_f^-(x).$$

We have to check that φ_f is negative on $L \setminus (R \cup O)$, positive on $R \setminus (L \cup O)$, 0 on $O \setminus (R \cup L)$, and non-zero on $(L \cup R) \setminus O$, and . First, we will show that φ_f is negative on $\widetilde{L}$, positive on $\widetilde{R}$, and 0 on $\widetilde{O}$ actually. Let $x \in \widetilde{L}$, then it is not part of $\mathbb{R} \setminus U$. Furthermore, the set $\mathbb{R} \setminus U$ is closed and for closed sets A the set distance function has the property that $d(x, A) = 0$ if and only if $x \in A$. This shows that $d(x, \mathbb{R} \setminus U) > 0$. By separatedness of $\widetilde{L}$ and $\widetilde{R}$ we also know that x cannot be an element of V, because U and V are necessarily disjoint. It follows that $d(x, \mathbb{R} \setminus V) = 0$. From this we conclude that $\varphi_f(x) < 0$ for $x \in \widetilde{L}$. Analogously, it can be shown that $\varphi_f(x) > 0$ for $x \in \widetilde{R}$.

If $x \in \widetilde{O}$, then $x \notin U \cup V$, which implies that $d(x, \mathbb{R} \setminus U) = d(x, \mathbb{R} \setminus V) = 0$. Precisely what is needed

Remark that $L \setminus (R \cup O) \subseteq \widetilde{L}$ and $R \setminus (L \cup O) \subseteq \widetilde{R}$, which tells us that φ_f is negative on $L \setminus (R \cup O)$ and positive on $R \setminus (L \cup O)$. Next, if $x \in L \cup R \setminus O$ it must be in either $\widetilde{L}$ or $\widetilde{R}$, as $L \cup R \cup O = \widetilde{L} \cup \widetilde{R} \cup \widetilde{O}$ and x cannot be in $\widetilde{O}$ From this we see that φ_f is non-zero on $L \cup R$. We conclude that this constructed φ_f is a continuous recourse sensitive attribution function.

Only if: We assume that we have a continuous recourse sensitive attribution function φ_f for f. Using this φ_f we can construct the required decomposition explicitly. Define

$$\widetilde{L} = \{x \in \mathcal{X} \mid \varphi(x) < 0, x \in L\},$$
$$\widetilde{R} = \{x \in \mathcal{X} \mid \varphi(x) > 0, x \in R\},$$
$$\widetilde{O} = \{x \in \mathcal{X} \mid \varphi(x) = 0, x \in O\}.$$

We see that $\widetilde{L} \cup \widetilde{R} \cup \widetilde{O} = \{\varphi_f \in \mathbb{R}\} \cap (L \cup R \cup O) = L \cup R \cup O$. We know that φ_f is continuous. This means that $\varphi_f(x) \leq 0$ on the closure of $\widetilde{L}$. It follows that $\mathrm{cl}(\widetilde{L}) \cap \widetilde{R} = \varnothing$, as $\varphi_f(x)$ is strictly positive on $\widetilde{R}$. Analoguesly , it can be argued that $\widetilde{L} \cap \mathrm{cl}(\widetilde{R}) = \varnothing$. Finally, $\varphi_f(x) = 0$ for all $x \in \widetilde{O}$. So, again $\varphi(x) = 0$ on the closure of $\widetilde{O}$. This guarantees that $\mathrm{cl}(\widetilde{O}) \cap \widetilde{L} = \mathrm{cl}(\widetilde{O}) \cap \widetilde{R} = \varnothing$, which verifies property 3 ■

Theorem 11 *Let $\delta > 0, \tau \in \mathbb{R}$, $f \colon \mathcal{X} \to \mathbb{R}$ and $C(x)$ be arbitrary, and let u_f be any utility function with $u_f(x, x) < \tau$ for all $x \in \mathcal{X}$. Then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exists a partition $\{\widetilde{K}_1, \widetilde{K}_2\}$ of $\widetilde{K}$ such that the sets*

$$\widetilde{L} = \Big(\bigcup_{i \in \widetilde{I}} L_i\Big) \cup \Big(\bigcup_{i \in \widetilde{K}_1} L_i\Big) \quad \textit{and} \quad \widetilde{R} = \Big(\bigcup_{j \in \widetilde{J}} R_j\Big) \cup \Big(\bigcup_{i \in \widetilde{K}_2} L_i\Big) \tag{5}$$

are separated.

Proof *If:* Suppose there exists a partition $\{\widetilde{K}_1, \widetilde{K}_2\}$ such that $\widetilde{L}$ and $\widetilde{R}$ are separated. Then existence of φ_f follows from Theorem 10: it is immediate that $\widetilde{L} \subseteq L$ and $\widetilde{R} \subseteq R$; and $\widetilde{L} \cup \widetilde{R} = L \cup R$ can be verified as follows: for any interval L_i that is contained both in $\mathcal{L}$ and in $\mathcal{R}$, we have $i \in \widetilde{K}$, so the interval is contained either in $\widetilde{L}$ or in $\widetilde{R}$. Any interval $L_i \in \mathcal{L}$ that is not in $\mathcal{R}$ is either contained in $\widetilde{L}$ or there exists $R_j \in \mathcal{R}$ such that $L_i \subset R_j$. In the latter case R_j must be contained in $\widetilde{R}$, because there cannot exist any $i^* \in \mathcal{I}$ such that $R_j \subseteq L_{i^*}$. If there were such an i^*, then we would have $L_i \subset R_j \subseteq L_{i^*}$, which would contradict the fact that all intervals in $\mathcal{L}$ are separated. By an analogous argument, any interval $R_j \in \mathcal{R}$ that is not in $\mathcal{L}$ is either contained in $\widetilde{R}$ or there exists $L_i \in \mathcal{L}$ that is contained in $\widetilde{L}$.

Only if: Suppose that $\widetilde{L}$ and $\widetilde{R}$ satisfy the conditions of Theorem 10. Then we will show that they must be of the form (5) for some partition $\{\widetilde{K}_1, \widetilde{K}_2\}$. To this end, we first observe that each interval $L_i \in \mathcal{L}$ must either be fully included in $\widetilde{L}$ or not included at all. Otherwise, the fact that $\widetilde{L} \cup \widetilde{R} = L \cup R$ would imply that part of the interval was included in $\widetilde{L}$ and the other part in $\widetilde{R}$, but then $\widetilde{L}$ and $\widetilde{R}$ would not be separated. Similarly, each interval $R_j \in \mathcal{R}$ must either be fully included in $\widetilde{R}$ or not included at all.

We can further restrict the intervals $L_i \in \mathcal{L}$ that can possibly be included in $\widetilde{L}$: if there exists some $R_j \in \mathcal{R}$ such that $L_i \subset R_j$, then $R_j \setminus L \neq \emptyset$ (otherwise L_i would not be a maximal interval), so R_j must be included in $\widetilde{R}$ to ensure that $\widetilde{L} \cup \widetilde{R} = L \cup R$. But then L_i cannot be included in $\widetilde{L}$, because otherwise $\widetilde{L}$ and $\widetilde{R}$ would not be separated. Similarly, no $R_j \in \mathcal{R}$ for which there exists some $L_i \in \mathcal{L}$ such that $R_j \subset L_i$, can be included in $\widetilde{R}$. This restricts attention to the intervals indexed by $\widetilde{I}$, $\widetilde{J}$ and $\widetilde{K}$.

We proceed to show that all intervals indexed by $\widetilde{I}$ and $\widetilde{J}$ must be included in $\widetilde{L}$ and $\widetilde{R}$, respectively. By symmetry, it is sufficient to show this for intervals L_i with $i \in \widetilde{I}$. For these, we have that $L_i \setminus R \neq \emptyset$ (otherwise R would contain an interval containing L_i), so that L_i must be included in $\widetilde{L}$ because $\widetilde{L} \cup \widetilde{R} = L \cup R$.

Finally, each interval indexed by $\widetilde{K}$ must be included either in $\widetilde{L}$ or in $\widetilde{R}$, but not in both, if we are to end up with separated sets $\widetilde{L}$ and $\widetilde{R}$ that satisfy $\widetilde{L} \cup \widetilde{R} = L \cup R$. Consequently, the intervals indexed by $\widetilde{K}$ should be partitioned among $\widetilde{L}$ and $\widetilde{R}$, as specified by the theorem. ■

Theorem 12 *Let $\delta > 0, \tau \in \mathbb{R}$, $C(x) = \{y \in \mathcal{X} \mid \|x - y\|_0 \leq 1\}$ and $f \colon \mathcal{X} \to \mathbb{R}$ be arbitrary. Then there exists a continuous recourse sensitive attribution function φ_f for f if and only if there exist $\widetilde{L}^i \subseteq L^i$ and $\widetilde{R}^i \subseteq R^i$ for all $i = 1, \ldots, d$ and $\widetilde{O} \subseteq O$ such that*

1. $\widetilde{O} \cup \bigcup_{i=1}^d (\widetilde{L}^i \cup \widetilde{R}^i) = O \cup \bigcup_{i=1}^d (L^i \cup R^i)$;
2. *All sets in $\{\widetilde{L}^1, \widetilde{R}^1, \ldots, \widetilde{L}^d, \widetilde{R}^d\}$ are pairwise separated;*
3. $\mathrm{cl}(\widetilde{O}) \cap \widetilde{L}^i = \varnothing$ *and* $\mathrm{cl}(\widetilde{O}) \cap \widetilde{R}^i = \varnothing$ *for all* $i = 1, \ldots, d$.

Proof Before we start proving both implications, we make the following observation. That is, the attribution φ_f is only allowed to be non-zero in the i'th component on the sets $\widetilde{L}^i$ and $\widetilde{R}^i$. Indeed, recourse sensitivity of φ_f tells us that $\varphi_f(x) = \gamma(y - x)$ for some $\gamma > 0$ and $\|x - y\| \leq \delta$, but most importantly y has to be of the form $y = x \pm \alpha e_i$ by the constraining set $C(x)$. The attribution is seen to be $\varphi_f(x) = \pm\gamma\alpha e_i$ and $\varphi_f(x)$ is only allowed to be non-zero in the i 'th component. By continuity of φ_f the above argument also extends to the closures of $\widetilde{L}^i$ and $\widetilde{R}^i$.

If: Just as in the one-dimensional case, we are able to construct a recourse sensitive function explicitly, using the set distance function $d(x, A)$ defined in (6). For each $\widetilde{L}^i$ and $\widetilde{R}^i$ find an open neighborhood $\widetilde{L}^i \subseteq U^i$ and $\widetilde{R}^i \subset V^i$ that is disjoint from all the other neighborhoods and $\widetilde{O}$. This is possible because of the pairwise separatedness. To see this, take one $\widetilde{L}^i$ and enumerate all the other $\widetilde{L}^j$ and $\widetilde{R}^j$ from $k = 1$ to $k = 2d - 1$ and denote them by $\widetilde{W}_k$. By the pairwise separatedness we can find open neighborhoods U_k^i for $\widetilde{L}^i$ and V_k for $\widetilde{W}_k$ that are disjoint. We can also find an open neighbourhood U_{2d}^i of $\widetilde{L}^i$ that is disjoint of $\widetilde{O}$, Then, take $U^i = \cap_{k=1}^{2d} U_k^i$. This is still an open set, as it is the finite intersection of open sets, and $\widetilde{L}^i \subseteq U^i$, because $\widetilde{L}^i$ is a subset of each of the U_k^i. The set U^i is also smaller than any of its components in the intersection, meaning that U^i is disjoint of all the other open neighborhoods. Repeat this procedure for every $\widetilde{L}^i$ and $\widetilde{R}^i$ to get our required op neighborhoods.

We are now ready to define φ_f. For each component set

$$\varphi_f^-(x)_i = \frac{d(x, \mathbb{R}^d \setminus U^i)}{1 + d(x, \mathbb{R}^d \setminus U^i)},$$
$$\varphi_f^+(x)_i = \frac{d(x, \mathbb{R}^d \setminus V^i)}{1 + d(x, \mathbb{R}^d \setminus V^i)},$$
$$\varphi_f(x)_i = \varphi_f^+(x)_i - \varphi_f^-(x)_i.$$

The attribution φ_f now becomes

$$\varphi_f(x) = \begin{bmatrix} \varphi_f(x)_1 \\ \varphi_f(x)_2 \\ \vdots \\ \varphi_f(x)_d \end{bmatrix}.$$

All the components of φ_f consist of continuous functions. So, φ_f is itself continuous. Next, note that if $x \in \widetilde{L}^i$ or $x \in \widetilde{R}^i$ for some i, it is also contained in U^i or V^i respectively. Because all U^i and V^i are mutually disjoint, we see that only $d(x, \mathbb{R}^d \setminus U^i)$ or $d(x, \mathbb{R}^d \setminus V^i)$ is non-zero. This ensure that only the i 'th component is non-zero, which is required be the remark at the start of this proof. Finally, if $x \in \widetilde{L}^i$, then $x \notin \widetilde{R}^i$ and $\varphi_f(x)_i < 0$, because $x \in U^i$ and $\mathbb{R}^d \setminus U^i$ is closed. Alternatively, if $x \in \widetilde{R}^i$, then $x \notin \widetilde{L}^i$ and $\varphi_f(x)_i > 0$, as is required.

For notational sake denote $L = \bigcup_{j=1}^d L^j$ and $R = \bigcup_{j=1}^d R^j$. To conclude, we note that

$$L^i \setminus \left(\bigcup_{\substack{j=1 \\ j \neq i}}^d L^j \cup R \cup O \right) \subseteq \widetilde{L}^i,$$
$$R^i \setminus \left(L \cup \bigcup_{\substack{j=1 \\ j \neq i}}^d R^j \cup O \right) \subseteq \widetilde{R}^i.$$

and

$$O \setminus (L \cup R) \subseteq \widetilde{O}.$$

Combining this with the argument above, we see that φ_f points in the correct directions on these sets. Furthermore, φ_f is also never zero on $(L \cup R) \setminus O$. By a similar reason as in the one dimensional case we see that $x \in (L \cup R) \setminus O$, implies that $x \in \widetilde{L}^i$ or $\widetilde{R}^i$ for some $i = 1, \ldots, d$. This implies $\varphi_f(x) \neq 0$. Finally, if $x \in \widetilde{O}$, then $x \notin U^i \cup V^i$ for all i. This immediately gives that $\varphi_f(x) = 0$, which shows that φ_f is 0 on $O \setminus (L \cup R)$. All together, we conclude that φ_f is a continuous recourse sensitive attribution function for f.

Only if: Assuming that φ_f is a recourse sensitive and continuous attribution function for f, define for all $i = 1, \ldots, d$ the sets

$$\widetilde{L}^i = \{x \in \mathcal{X} \mid \varphi_f^i(x) < 0, x \in L^i\},$$
$$\widetilde{R}^i = \{x \in \mathcal{X} \mid \varphi_f^i(x) > 0, x \in R^i\},$$
$$\widetilde{O} = \{x \in \mathcal{X} \mid \varphi_f(x) = 0, x \in O\}.$$

These sets form the required partition, because

$$\widetilde{O} \cup \bigcup_{i=1}^{d} \widetilde{L}^i \cup \widetilde{R}^i = \bigcup_{i=1}^{d} \{\varphi_f^i \in \mathbb{R}\} \cap (L^i \cup R^i \cup O) = O \cup \bigcup_{i=1}^{d} L^i \cup R^i,$$

We can now verify properties (2) and (3) by using the continuity of φ_f. Note that $\varphi_f^i(x) < 0$, implies that only the i'th component can be non-zero and that $x \in \widetilde{L}^i$, by the remark at the start of the proof. By continuity of φ_f it follows that $\varphi_f^i(x) \leq 0$ and $\varphi_f^j(x) = 0$ for all $x \in \mathrm{cl}(\widetilde{L}^i)$. On all the other $\widetilde{L}^j$ or $\widetilde{R}^j$ it must be that $\varphi_f^j(x)$ is strictly non-zero, or positive for $\varphi_f^i(x)$ and $\widetilde{R}^i$. We see that $\mathrm{cl}(\widetilde{L}^i)$ is disjoint from all other $\widetilde{L}^j$ or $\widetilde{R}^j$. This argument holds for all i and we can proof it analogously for $\widetilde{R}^i$. This verifies property (2).

Finally, $\varphi_f(x) = 0$ for all $x \in \widetilde{O}$. So, again $\varphi_f(x) = 0$ on $\mathrm{cl}(\widetilde{O})$. The function φ_f will be non-zero on each of the sets $\widetilde{L}^i$ and $\widetilde{R}^i$. Thus, $\mathrm{cl}(\widetilde{O}) \cap \widetilde{L}^i = \mathrm{cl}(\widetilde{O}) \cap \widetilde{R}^i = \varnothing$ for all $i = 1, \ldots, d$. This verifies property (3). ■

Appendix E. Additional Details for Section 4

In Section 4, it is mentioned that recourse can be provided when the model is very simply, for example when using a linear classifier. This is also noted by Ustun et al. (2019). In this section we will expand on this statement. We will also give an example of a classifier f that is non-linear, but does allow a linear representation $f(x) = \beta^\top g(x)$ using higher order or more abstract features. In this example, the features $g(x)$ are still interpretable and providing a continuous recourse sensitive attribution function in terms of the features $g(x)$ is possible.

E.1 Linear Classifiers Admit Recourse

Consider the binary classification task using $f(x) = \beta^\top x$ for some vector β. Recall that the utility function is given by $u_f(x, y) = f(y) \geq 0$. A point is classified as the negative class if $f(x) < 0$ and as the preferred class if $f(x) \geq 0$. In light of Theorem 6 we see that U is given by

$$U = \{x \in \mathbb{R}^d \mid \beta^\top x \geq 0\},$$

which is a convex and closed set. Using Theorem 6 we conclude that a recourse sensitive and robust attribution function exists.

E.2 Attribution for Abstract Features

Consider the non-linear classifier $f(x) = \|x\|^2 - 1$, which classifies if a point is inside the circle or outside the circle. To show that there are no continuous and recourse sensitive functions for this classifier we consider the following two cases:

1. $\delta > 0$ and $C(x) = \{y \in \mathbb{R}^2 \mid \|x - y\|_0 \leq 1\}$;
2. $1 \leq \delta < 2$ and $C(x) = \mathbb{R}^2$.

First, we will show the single feature case, because it follows from Theorem 12. The second case requires special arguments and will follow afterwards.

E.2.1 $\delta > 0$ AND $C(x) = \{y \in \mathbb{R}^2 \mid \|x - y\|_0 \leq 1\}$

To apply Theorem 12 we first find all 4 sets L^1, R^1, L^2 and R^2. If we know L^1, then we can find all the other sets as well by the symmetry of f. The set L^1 consists of all points such that you cross the decision boundary when you subtract $[\delta, 0]^\top$ from the input point. This is the strip to right of the circle with width δ and all the points within the circle that also do not lie in the translated circle $D_1(0) + [\delta, 0]^\top$, where $D_1(0) = \{y \in \mathbb{R}^2 \mid \|y\| < \delta\}$. In set notation

$$\begin{aligned} L^1 &= \left\{ \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \mid \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \alpha \in (0, \delta) \right\} \cup \left(D_1(0) \setminus \left(D_1(0) + \begin{bmatrix} \delta \\ 0 \end{bmatrix} \right) \right) \\ &=: L^1_{\text{out}} \cup L^1_{\text{in}}. \end{aligned}$$

Note that L^1_{out} and L^1_{in} are two disjoint connected components. The set R^1 can be given in a similar form, with the α replaced by $-\alpha$ and the vector $[\delta, 0]^\top$ with $[-\delta, 0]^\top$. The sets L^2 and R^2 can be obtained by rotating the sets L^1 and R^1 with $\frac{\pi}{2}$.

Take any $\alpha \in (0, \delta)$ and consider the point $x = [1 + \alpha, 0]^\top$. This point is only an element of L^1 and not of any of the other sets. As x is contained in L^1_{out} and L^1_{out} is connected, we know that it must be that $L^1_{\text{out}} \subseteq \widetilde{L}^1$ for any decomposition. Similarly, we see that $L^2_{\text{out}} \subseteq \widetilde{L}^2$. However, L^1_{out} and L^2_{out} are not disjoint, because $\sqrt{\alpha}[1/\sqrt{2}, 1/\sqrt{2}]$ is an element of both sets for $\alpha \in (1, \min(2, 2\delta))$. It follows that $\widetilde{L}^1$ and $\widetilde{L}^2$ cannot be separated and Theorem 12 tells us that no continuous single feature attribution function can exist.

E.2.2 $1 \leq \delta < 2$ AND $C(x) = \mathbb{R}^2$

Note that in all cases that follow an attribution can be given for the region outside of the circle by

$$\varphi_f(x) = \begin{bmatrix} -x_1 f(x) \\ -x_2 f(x) \end{bmatrix}.$$

So, we only have to focus on the region inside the circle. If $0 < \delta < 1$, then we cannot cross the decision boundary for any $x \in D_{1-\delta}(0)$. In that region any value for the attribution is allowed and extending the above φ_f to the whole plane gives us a valid continuous recourse sensitive attribution function.

When $\delta > 2$, we can cross the decision boundary for any $x \in D_1(0)$ by moving in any direction with length δ. So, inside the circle we could set $\varphi_f(x)$ to any direction. A full continuous recourse

sensitive attribution function would be given by

$$\varphi_f(x) = \begin{cases} \begin{bmatrix} -x_1 f(x) \\ -x_2 f(x) \end{bmatrix} & \|x\| \geq 1 \\ \begin{bmatrix} f(x) \\ 0 \end{bmatrix} & \|x\| < 1 \end{cases}.$$

Now, we can discuss the case when $1 \leq \delta < 2$. Again, the attribution outside of the circle does not pose a problem. Inside the circle we can identify two regions. Within $D_{\delta-1}(0)$ we can move in any direction of length δ to cross the decision boundary. Indeed, take an $x \in D_{\delta-1}(0)$ and note that the worst direction to cross the decision boundary is $-x$. We can scale this vector with $\frac{\delta}{\|x\|}$ to get a vector of length δ. Using $\|x\| < \delta - 1$, we see that moving in that direction crosses the decision boundary, as

$$\|x - \frac{\delta}{\|x\|}x\| = \left|1 - \frac{\delta}{\|x\|}\right| \|x\| = |\delta - \|x\|| > 1.$$

In the strip with $\delta - 1 \leq \|x\| < 1$, the set of feasible direction is more complicated. However, the important observation is that $-x$ is not contained in it. So, for any attribution φ_f it cannot be that $\varphi_f(x) = -\alpha x$ for any $\alpha > 0$. To conclude that φ_f has a zero we will use the following Lemma, which can be seen as a generalization of the intermediate value theorem.

Lemma 18 (Poincaré-Bohl) *Assume that U is an open bounded neighborhood of $\mathbb{R}^d$, with $0 \in U$, and that $f\colon \operatorname{cl}(U) \to \mathbb{R}^d$ is a continuous function such that*

$$f(x) \notin \{\alpha x \colon \alpha > 0\}, \textit{for every } x \in \operatorname{cl}(U) \setminus U.$$

Then, there is an $x_0 \in \operatorname{cl}(U)$ such that $f(x_0) = 0$.

Proof See Theorem 2 in Fonda and Gidoni (2016). ■

Applying Lemma 18 to the function $-\varphi_f(x)$ immediately gives that there is some $x \in D_1(0)$ such that $-\varphi_f(x) = 0 \iff \varphi_f(x) = 0$, which is not allowed if φ by recourse sensitivity.

However, if we write this function as a linear function of a feature map consisting of linear and quadratic terms. The feature map g and coefficients are given by

$$g(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}, \qquad \beta = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}.$$

The function f is then represented by

$$f(x) = \beta^\top g(x),$$

and we could provide recourse by communicating

$$\varphi_f(x) = \begin{bmatrix} -x_1 f(x) \\ -x_2 f(x) \\ -f(x) \\ -f(x) \\ 0 \end{bmatrix}.$$

This attribution will only be 0 on the decision boundary, which is allowed, and in almost all other cases the first two components will point towards the decision boundary. The first two components are only zero when $x = 0$. In that case the first two components do not point towards the decision boundary, but the final two components do provide information on which (higher-level) action has to be taken to change the class. Namely, it tells the user to increase the norm, in whatever way possible.

The above argument shows that, if it is possible to write the function f as some linear function $f(x) = \beta^\top g(x)$, it will be possible to provide recourse in terms of the higher level features of $g(x)$.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in neural information processing systems, NeurIPS*, 2018.

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, and Himabindu Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2021.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 Workshop on Human interpretability in Machine Learning*. ICML, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 2020.

Georgios Arvanitidis, Soren Hauberg, and Bernhard Schölkopf. Geometrically enriched latent spaces. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, Proceedings of Machine Learning Research. PMLR, 2021.

Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis*. Springer Science & Business Media, 2009.

Claude Berge. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation, 1997.

Blair L. Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *arXiv preprint arXiv:2212.11870*, 2022.

Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. In *International Conference on Learning Representations, ICLR*, 2022.

Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Conference on Fairness, Accountability, and Transparency, FAccT*. ACM, 2022.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*. Springer, 2020.

Arun Das and Paul Rad. Opportunities and challenges in Explainable Artificial Intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems, NeurIPS*, 2018.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems, NeurIPS*, 2019.

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

Paul Erdös. Some remarks on the measurability of certain sets. *Bulletin of the American Mathematical Society*, 1945.

Andrea Ferrario and Michele Loi. The robustness of counterfactual explanations over time. *IEEE Access*, 2022.

Alessandro Fonda and Paolo Gidoni. Generalizing the poincaré–miranda theorem: the avoiding cones condition. *Annali di Matematica Pura ed Applicata*, 2016.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. Counterfactual evaluation for explainable ai. *arXiv preprint arXiv:2109.01962*, 2021.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Conference on Artificial Intelligence, AAAI*. AAAI Press, 2019.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 2018.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2023.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in neural information processing systems, NeurIPS*, 2019.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*. Association for Computational Linguistics, 2020.

Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. Diagnosing ai explanation methods with folk concepts of behavior. In *Conference on Fairness, Accountability, and Transparency, FAccT*. ACM, 2023.

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

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

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, 2021.

Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2021.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science. Springer, 2019.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2019.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 2020.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing systems, NeurIPS*, 2017.

Bert Mendelson. *Introduction to topology*. Courier Corporation, 1990.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

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

Martin Pawelczyk, Tobias Leemann, Asia Biega, and Gjergji Kasneci. On the trade-off between actionable explanations and the right to be forgotten. In *International Conference on Learning Representations, ICLR*, 2022.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Conference on AI, Ethics, and Society, AIES*. AAAI, ACM, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining, SIGKDD*. ACM, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science. Springer Nature, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations, ICLR*, Workshop Track Proceedings, 2014.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Conference on AI, Ethics, and Society, AIES*. AAAI, ACM, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Proceedings of the 2017 Workshop on Visualization for Deep Learning*. ICML, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research. PMLR, 2017.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *Advances in Neural Information Processing Systems, NeurIPS*, 2021.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Conference on Fairness, Accountability, and Transparency, FAccT*. ACM Press, 2019.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2011.

Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 2017.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 2021.