Attribution-based Explanations that Provide Recourse Cannot be Robust

Hidde Fokkema — University of Amsterdam, Korteweg de Vries Institute for Mathematics

Rianne de Heide — Vrije Universiteit Amsterdam

Tim van Erven — University of Amsterdam, Korteweg de Vries Institute for Mathematics

Summary

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

Utility

A **Utility function**, $u_f(x, y)$, measures the utility increase by a user when changing x to y. A user is satisfied if $u_f(x,y) \geq \tau$ for a **threshold** τ. Some examples,

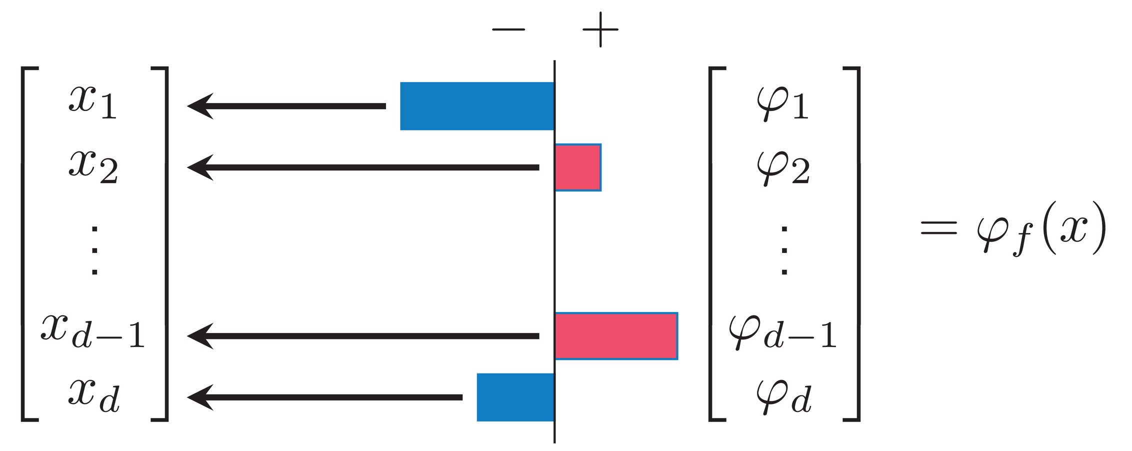
- **Flip the class label**:
  $u_f(x,y) = f(y) \geq 0$.
- **Increase score by amount** τ:
  $u_f(x,y) = f(y) - f(x) \geq \tau$.
- **Increase probability by** $p \times 100\%$:
  $u_f(x,y) = \dfrac{f(y)}{f(x)} \geq 1 + p$.

Attribution Methods

Machine learning model f, e.g. a classifier, and **Attribution function** φ_f are given by :

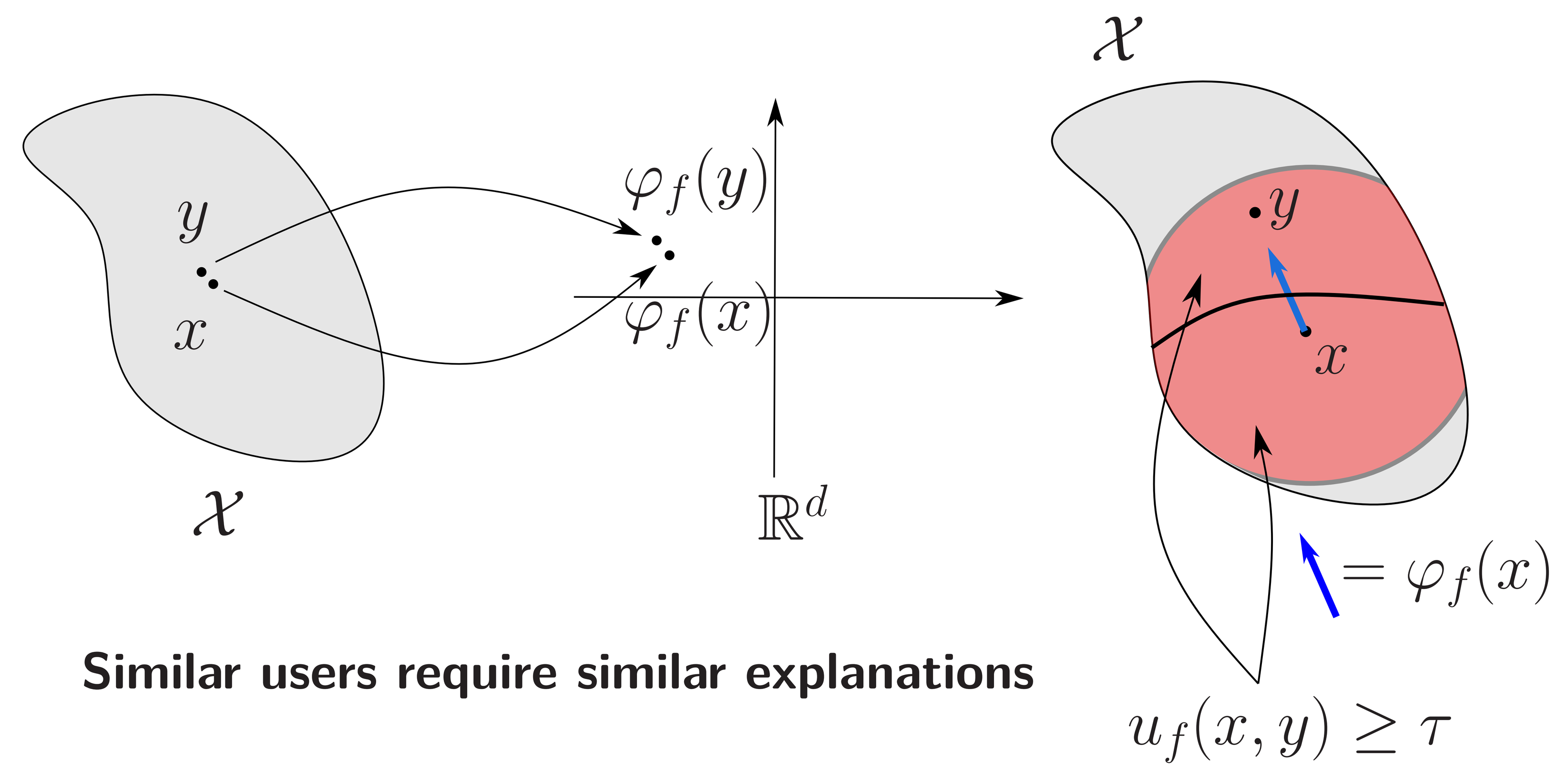
$$f \colon \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}, \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \mapsto y, \qquad \varphi_f \colon \mathcal{X} \to \mathbb{R}^d.$$

Indicating importance:



Robustness & Recourse Sensitivity

An attribution function φ_f for f is called **Robust** if it is continuous.



Informally, an attribution function is **Recourse Sensitive** if the user can achieve a sufficient utility increase when moving in the direction of $\varphi_f(x)$.

Impossibility Result

For Classification

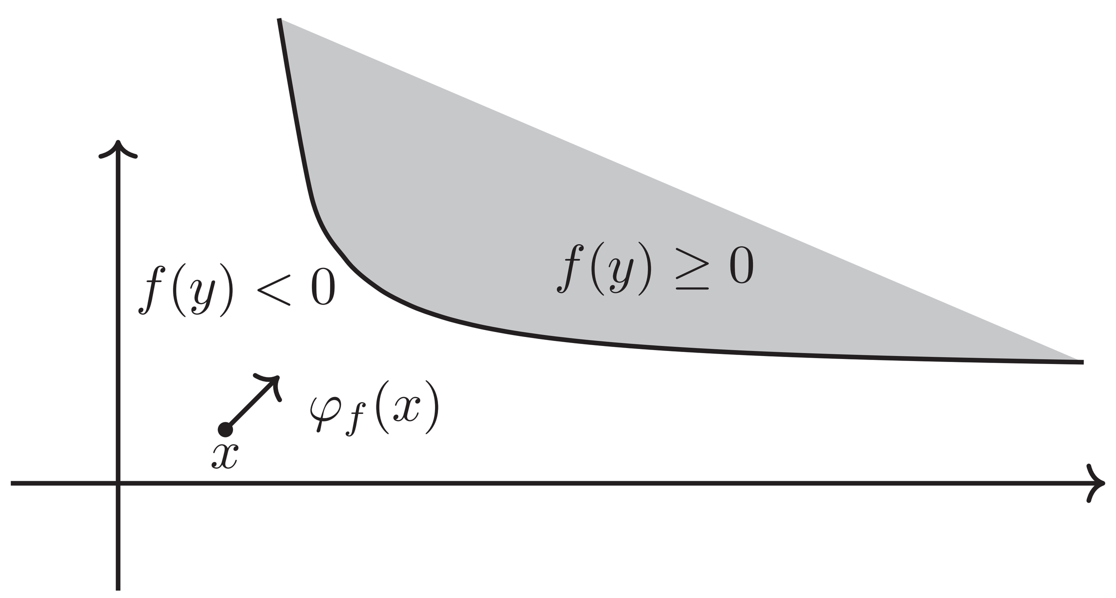
Theorem 1. *Suppose $\mathcal{X} = \mathbb{R}^d$, $u_f(x,y) = f(y)$, $\tau = 0$ and $\delta > 0$. Then, there exists a continuous classifier $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and continuous.*

General

Theorem 2. *If u_f is of the form $u_f(x,y) = \tilde{u}(f(x), f(y))$ and if there exist $z_1, z_2 \in \mathbb{R}^d$ such that $\tilde{u}(z_1, z_2) \geq \tau$ and $\tilde{u}(z_1, z_1) < \tau$. Then there exists a continuous $f \colon \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and robust.*

Sufficient Conditions for Recourse with Robustness

- **Theorem** 2 implies the existence of continuous functions f for which no attribution function can both provide recourse and be robust.
- But for specific **nice** f functions this may still be possible.
- Which functions are **nice** enough?

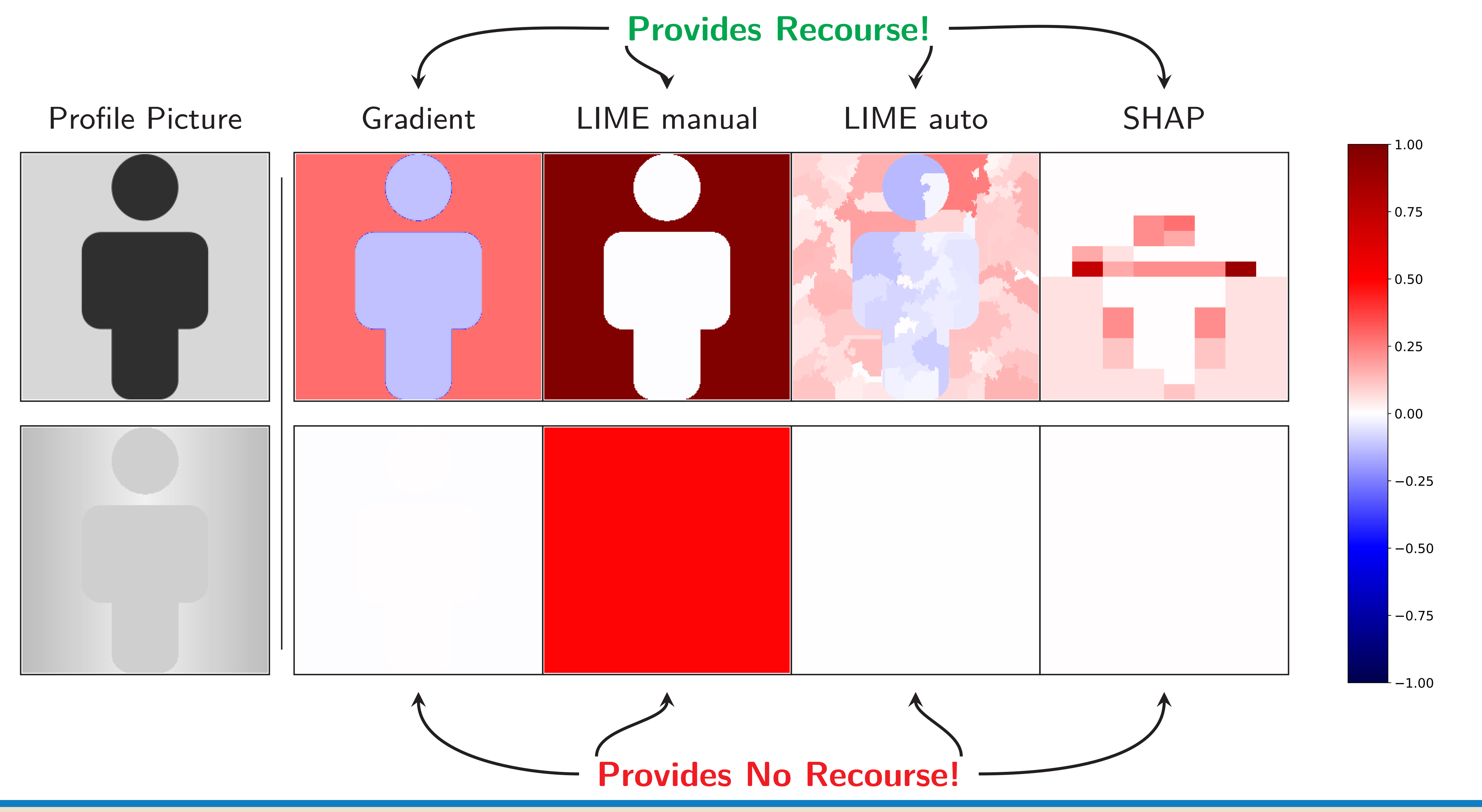


Theorem 3. *Let $u_f(x,y) = f(y)$ and $\tau = 0$, let $\delta > 0$ be arbitrary and take $f \colon \mathcal{X} \to \mathbb{R}$ to be any continuous function. If the set $U = \{y \in \mathcal{X} \mid f(y) \geq 0\}$ is convex, then the attribution method*

$$\varphi_f(x) := \operatorname*{argmin}_{y \in U} \|y - x\| - x = P_U(x) - x$$

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

Profile picture example



In the paper & Link

(1) Suggestions to circumvent impossibility.
(2) Generalisation of Sufficient Conditions.
(3) Fully characterising when Recourse and Robustness is possible in the 1-dimensional case.
(4) Proofs.

Read our paper online!