University of Amsterdam
Korteweg-De Vries instituut for Mathematics
VU Vrije Universiteit Amsterdam

Attribution-based Explanations that Provide Recourse Cannot be Robust

Joint work with
Rianne de Heide and Tim van Erven

2022-11-18

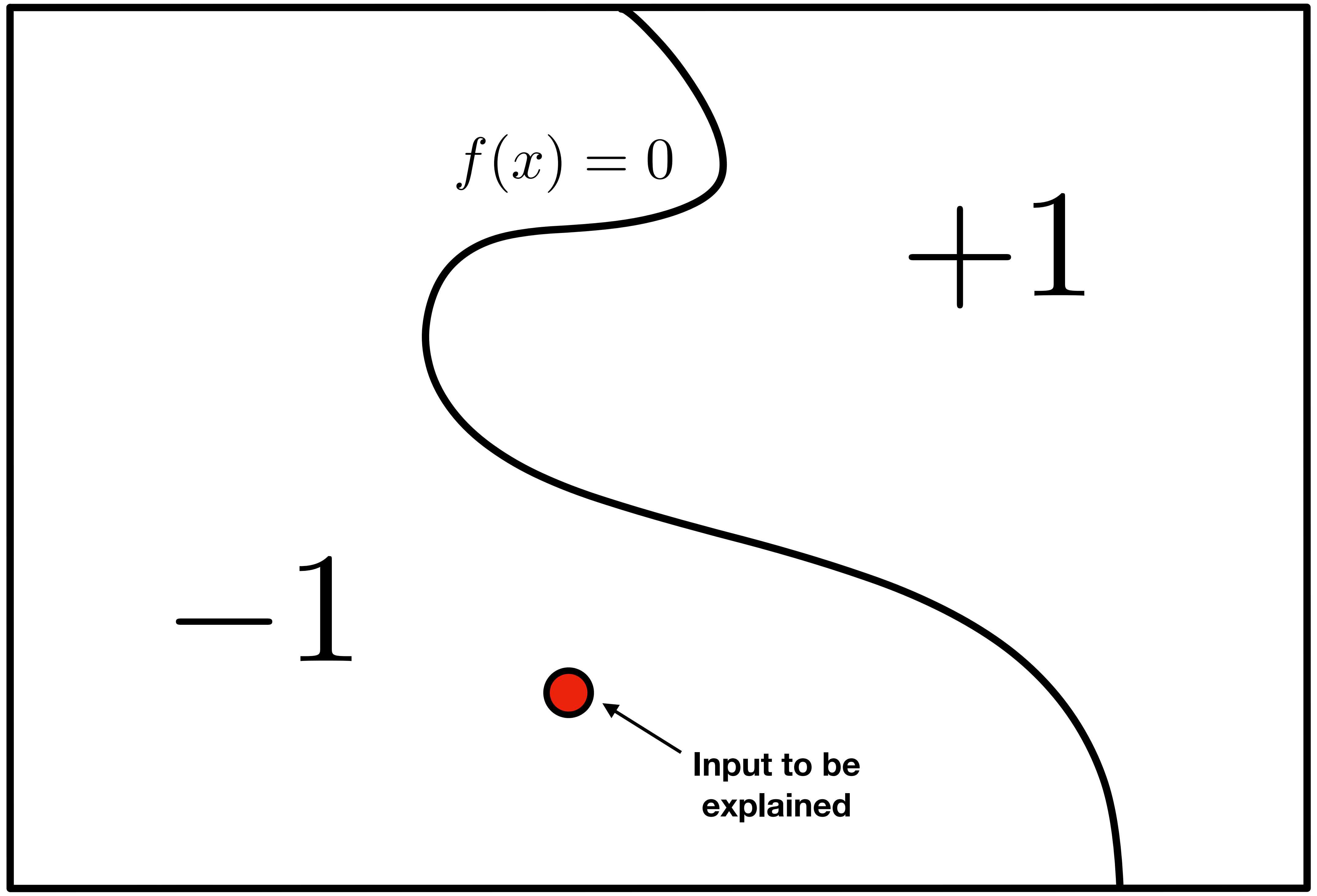
Programme of today

- Attribution Methods
- Recourse and Robustness
- Impossibility result
- When Recourse is possible

Attribution methods

Setting

Post-Hoc and local explanations



Machine learning model, e.g. a classifier:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to [0,1], \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mapsto y$$

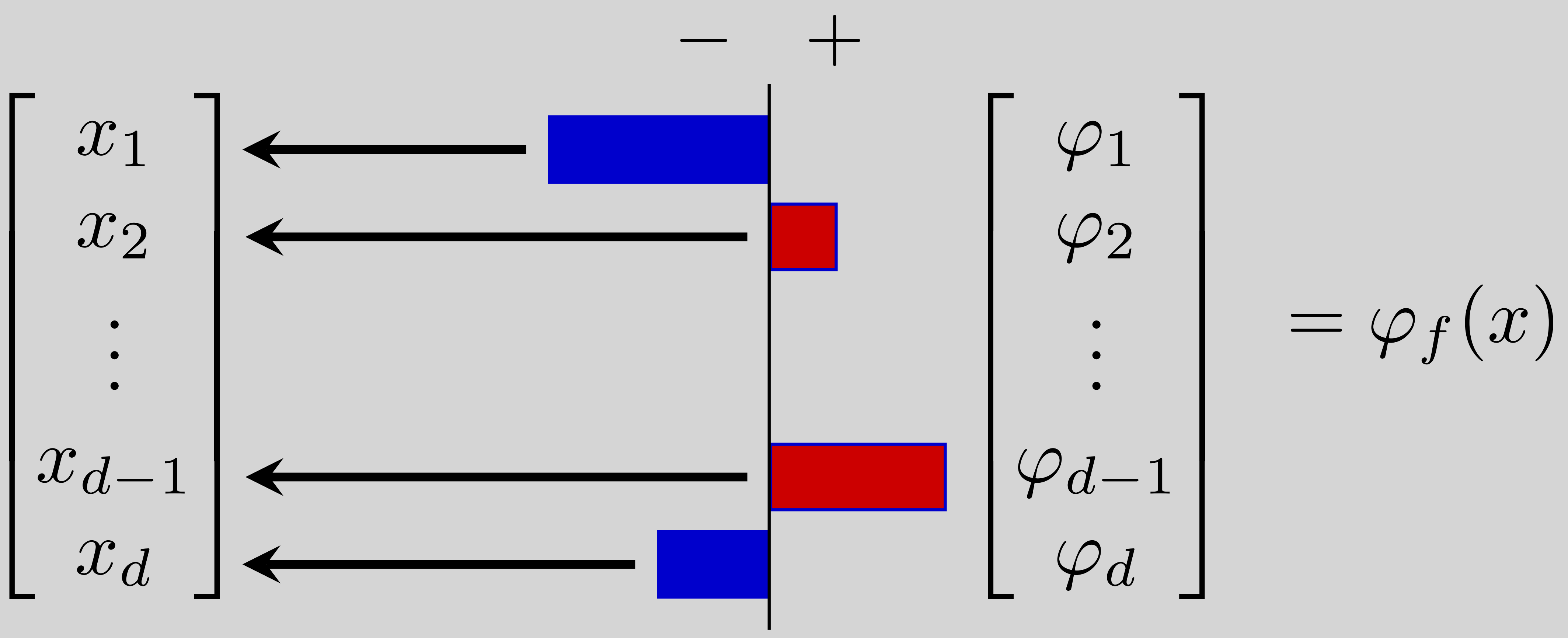
- **Local:** Only explain the part of f that is relevant for x
- **Post-Hoc:** The function f is given and fixed

Setting

Attribution methods

Machine learning model, e.g. a classifier:

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to [0,1], \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mapsto y$$



$\varphi_f(x) \in \mathbb{R}^d$ attributes a weight to each feature which explains how important the feature was for the classification of x of f

Counterfactuals as attributions

Definition

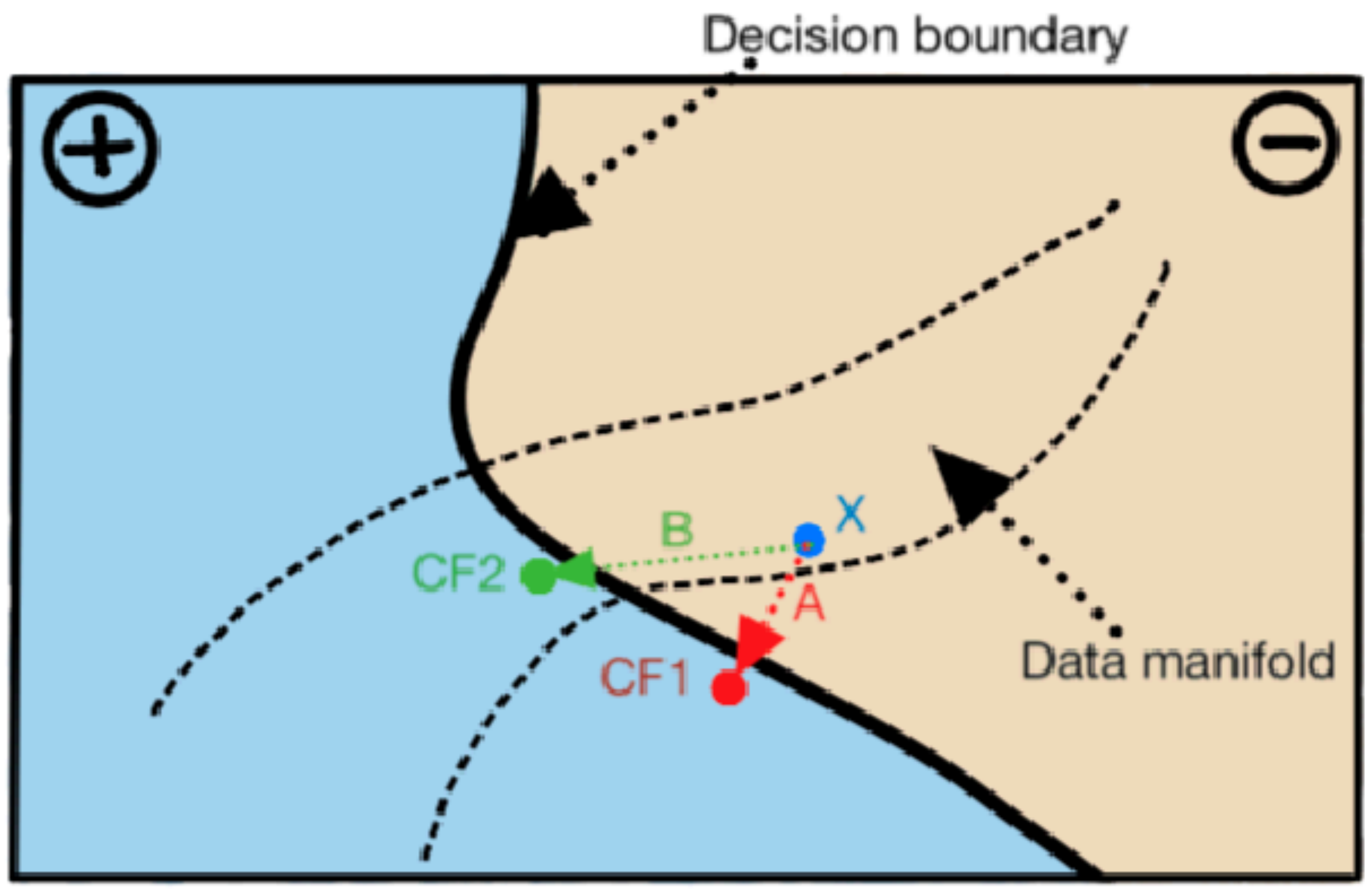
Consider Binary classification $f\colon \mathcal{X} \to \{-1,1\}$ and
let $x \in \mathcal{X}$.

A ***counterfactual*** x^{CF} for x is

$$x^{\mathrm{CF}} \in \arg\min_{y \in C} \|x - y\| \quad \mathrm{s.t.} \quad f(x^{\mathrm{CF}}) \neq f(x)$$

Counterfactuals can be seen as Attributions.
Write

$$\varphi_f(x) = x^{\mathrm{cf}} - x$$



[5]Image source: [Verma et al., 2020]

What are Good Explanations/Attributions?

- How to say some explanations are better than others?
- What is the (implicit) goal of the explanations?

[3] Image source: [Verma et al., 2020]

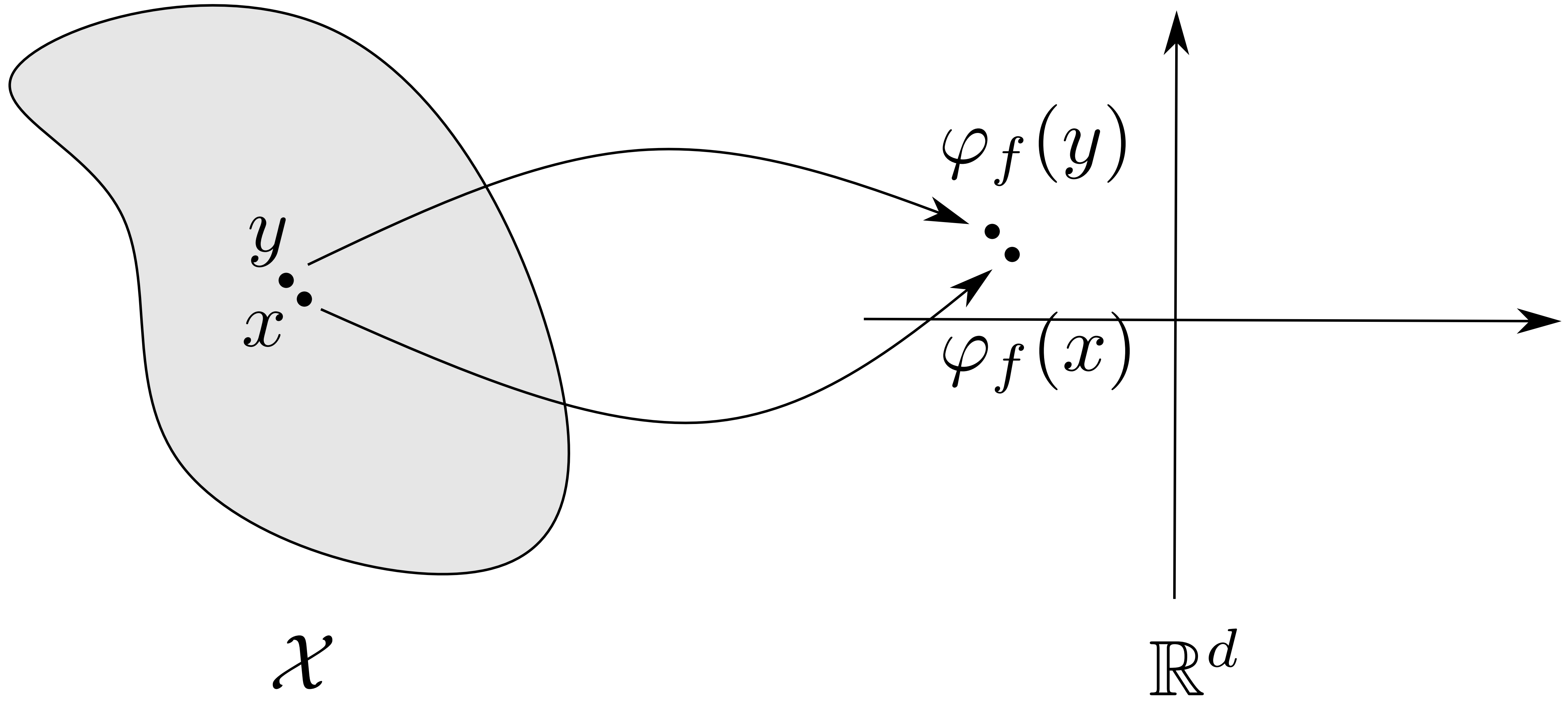
Robustness & Recourse sensitivity

Robustness

Definition

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

Similar users require similar explanations



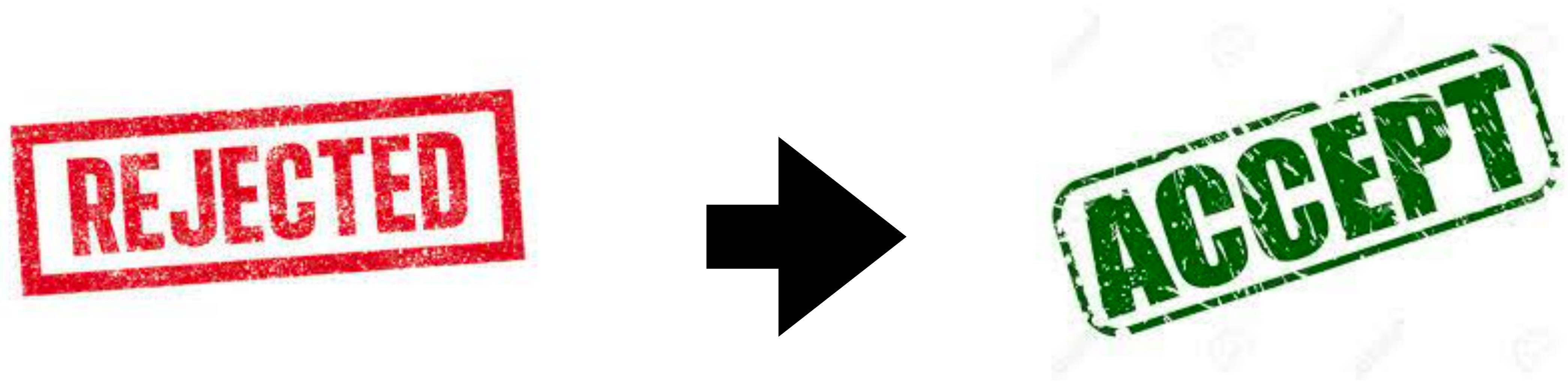
Recourse sensitivity

Motivation

User has some goal in mind:

- Wants to get a loan
- Increase their credit score
- Increase a probability
- Wants to upload a profile picture to get an OV card.

The explanation should allow the user to reach this goal

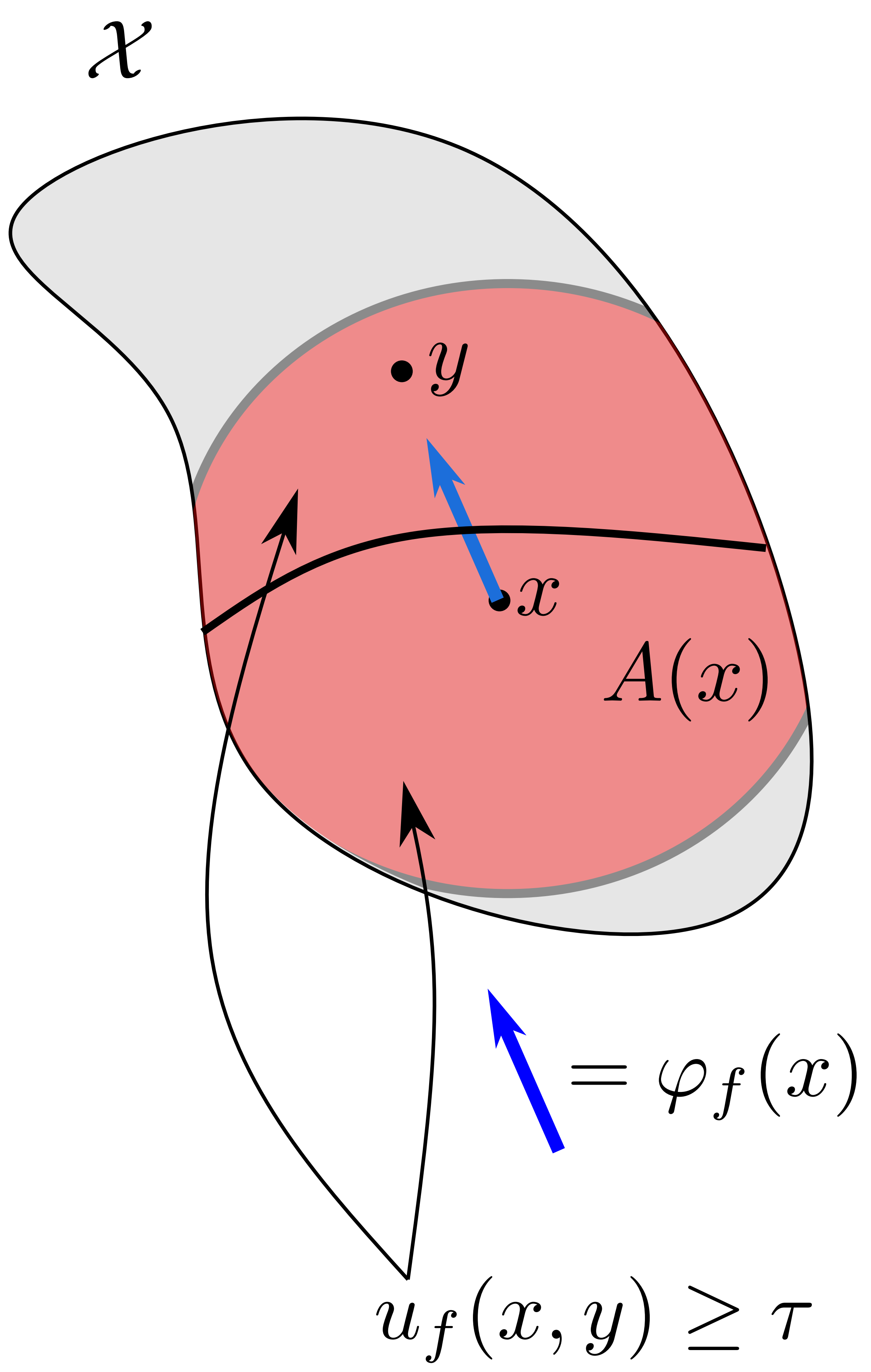


Recourse sensitivity

Informal definition

An Attribution method is called ***Recourse Sensitive*** if the user can achieve a sufficient utility increase when moving in the direction of $\varphi_f(x)$

This is very weak form of Recourse!



Recourse sensitivity

Definition

Definition

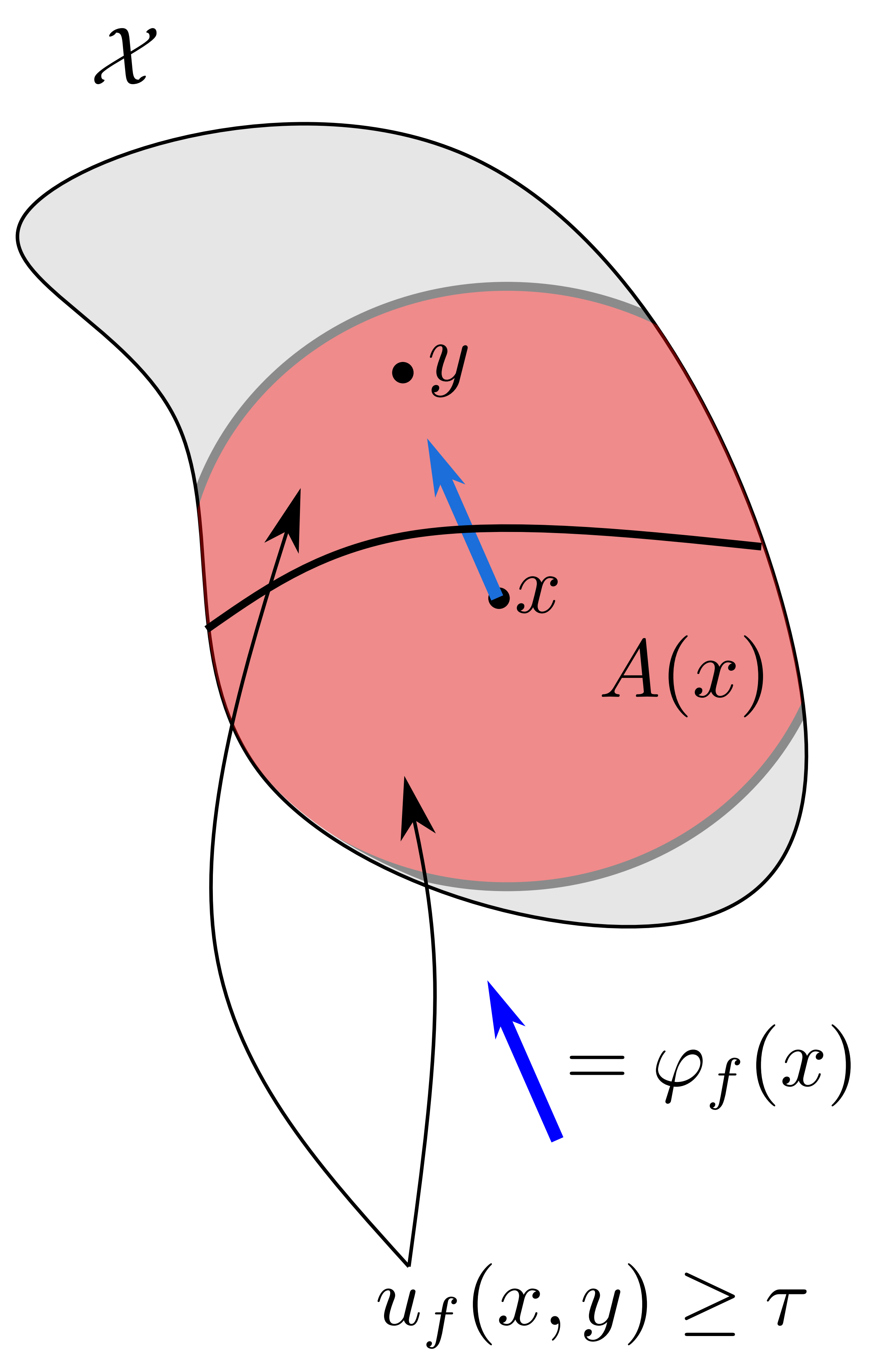
Consider the points close to x that achieve sufficient utility

$$U(x) = \{y \in \mathcal{X} \mid u_f(x, y) \geq \tau,\ \|x - y\| \leq \delta\}$$

An Attribution function φ_f is called ***Recourse Sensitive*** if

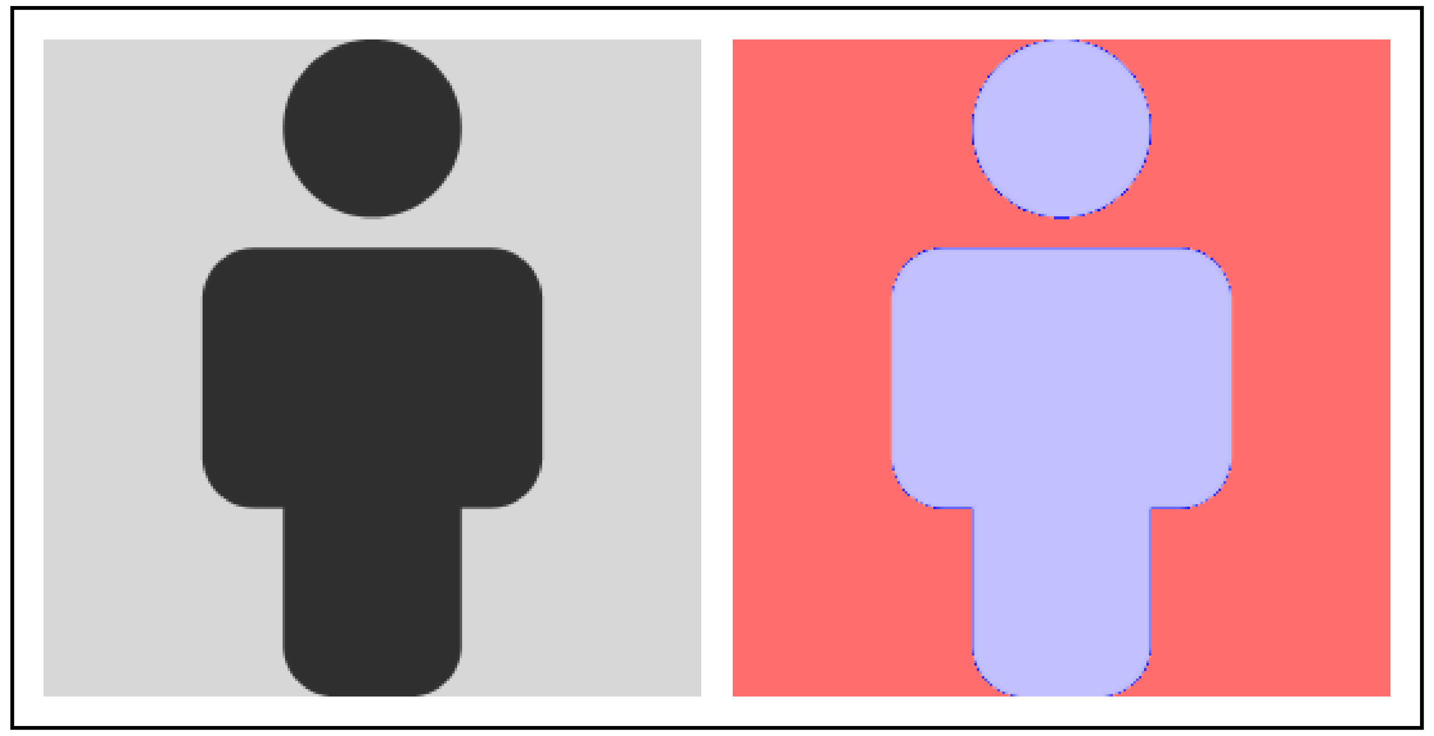
$$\varphi_f(x) = \alpha(y - x), \qquad \alpha > 0 \text{ and } y \in U(x),$$

for all $x \in \mathcal{X}$ for which $U(x) = \emptyset$.

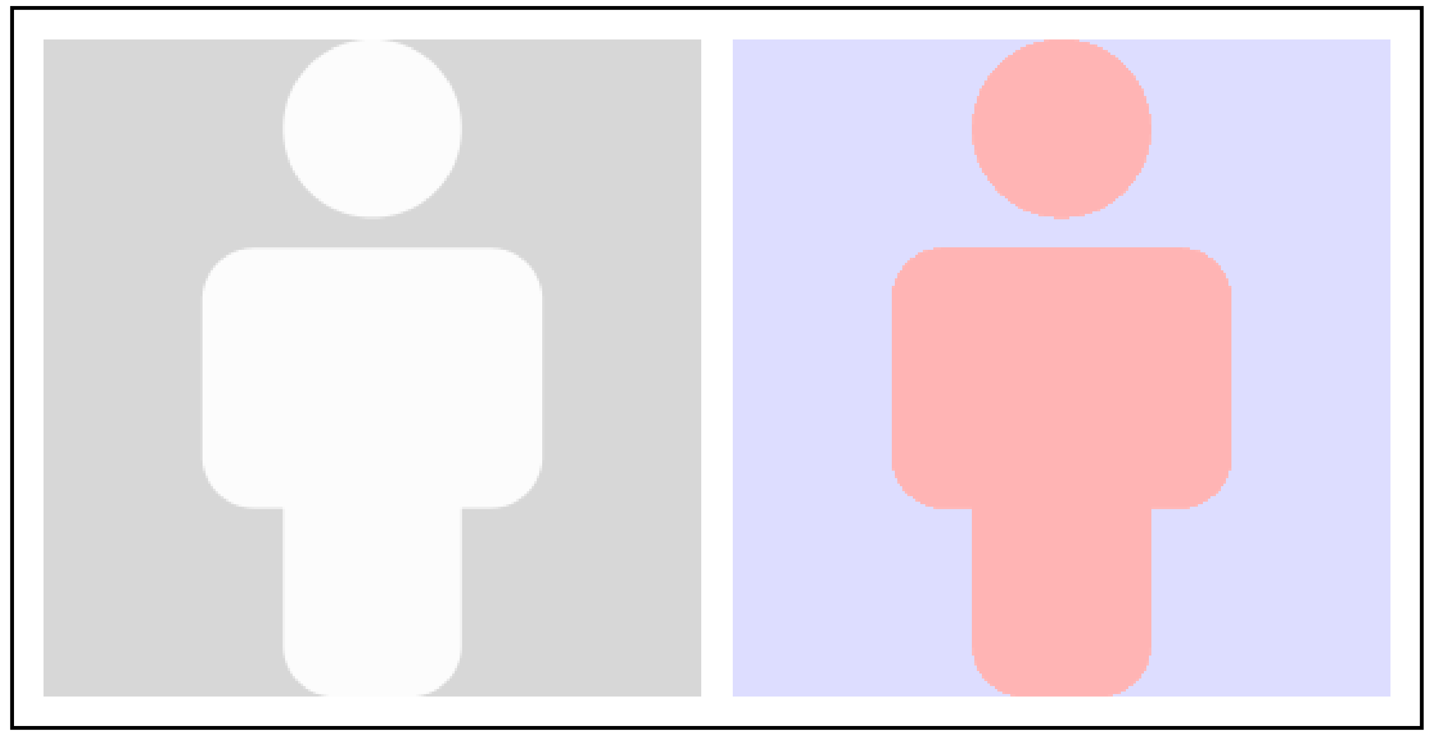


Recourse sensitivity

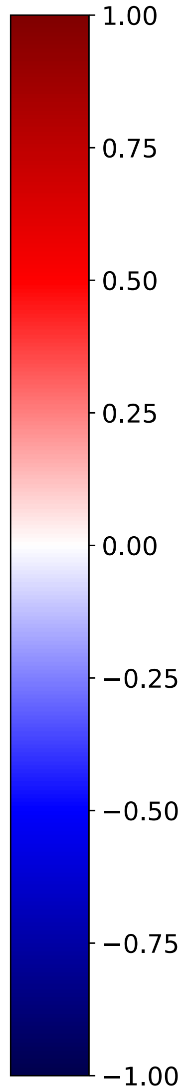
Example



(a) Accepted profile picture



(b) Rejected profile picture



Recourse sensitivity

Utility

Measure if some utility $u_f \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exceeds some threshold $u_f(x, y) \geq \tau$:

- Preferable class: $u_f(x, y) = f(y) \geq 0$
- Increase score: $u_f(x, y) = f(y) - f(x) \geq \tau$
- Decrease a probability: $u_f(x, y) = \dfrac{f(x)}{f(y)} \geq \dfrac{1}{1 - p} = \tau$

Impossibility

Impossibility result

Attribution methods cannot always

- Provide Recourse
- Be Robust

Impossibility result

Specific case (Binary classification)

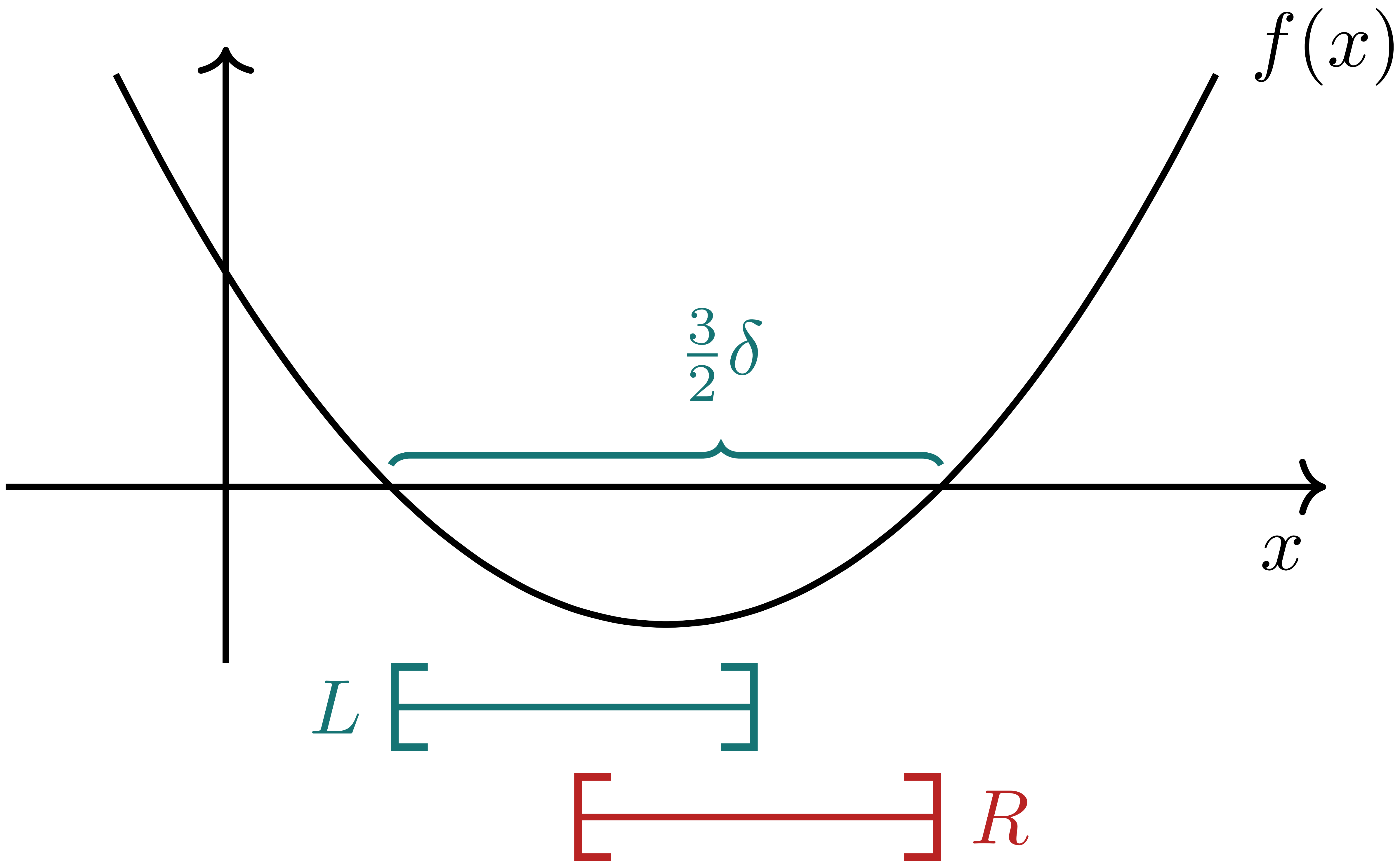
Setting

- $\mathscr{X} = \mathbb{R}^d$,
- $u_f(x, y) = f(y)$,
- $\tau = 0, \delta > 0$.

Theorem

There exists a continuous function f such that no attribution method φ_f can be both recourse sensitive and continuous

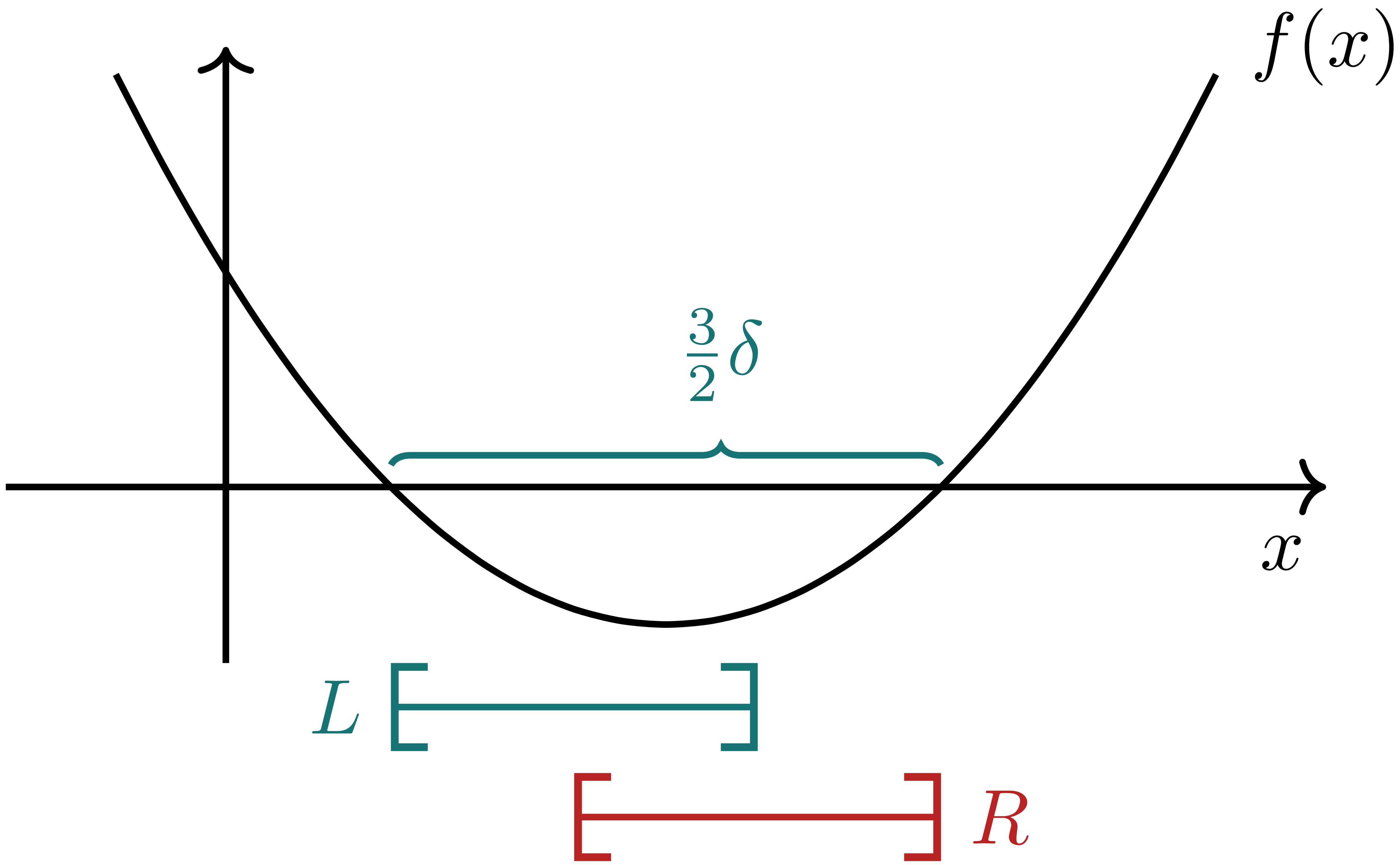
Proof sketch



$R = \{x \mid$ recourse is possible by moving at most δ left$\}$

$L = \{x \mid$ recourse is possible by moving at most δ left$\}$

Proof sketch

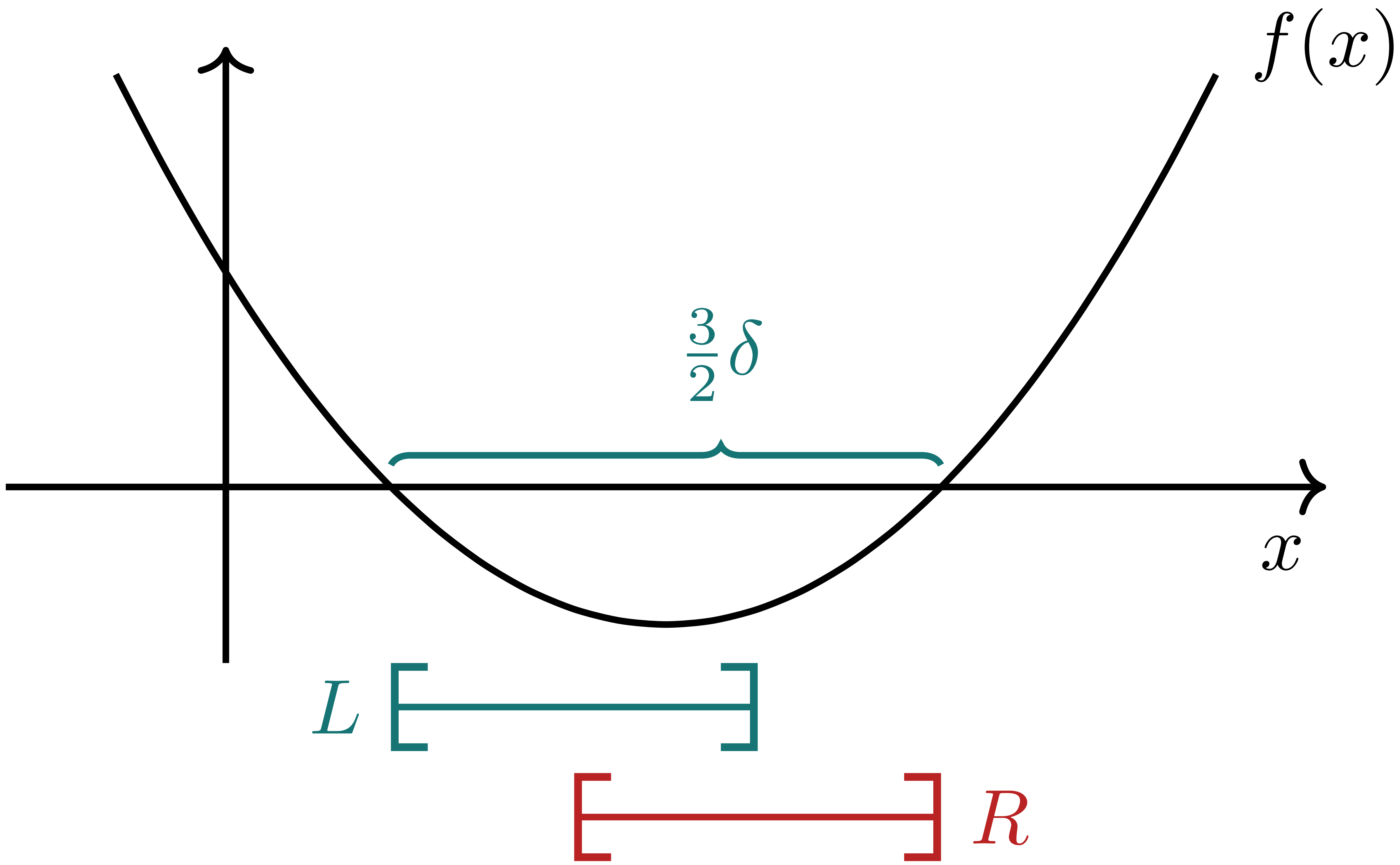


$R = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

$L = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

$$\varphi_f(x) = \begin{cases} < 0 & \text{for } x \in L \backslash R \\ > 0 & \text{for } x \in R \backslash L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

Proof sketch



$R = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$
$L = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

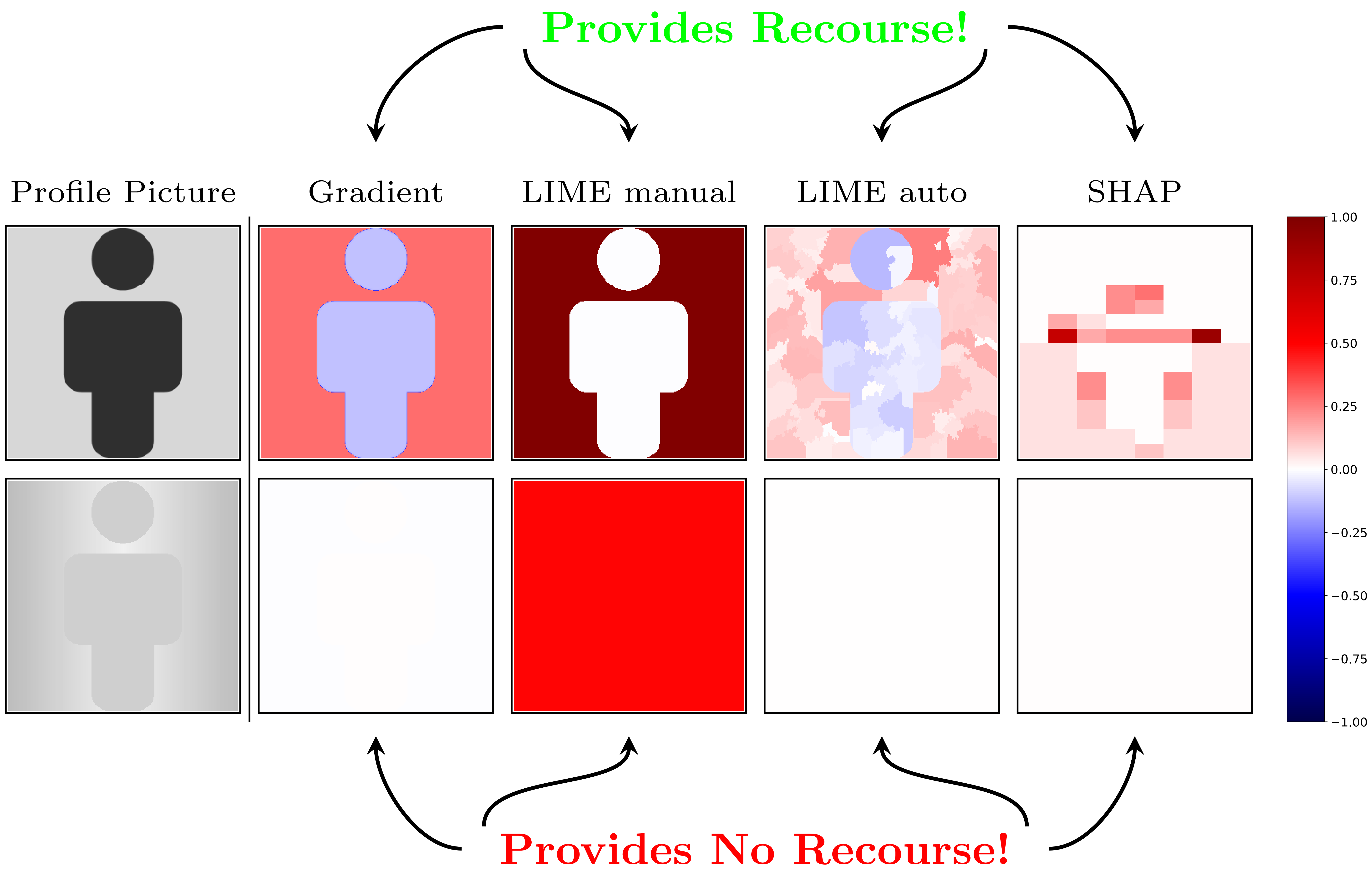
$$\varphi_f(x) = \begin{cases} < 0 & \text{for } x \in L \backslash R \\ > 0 & \text{for } x \in R \backslash L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

But this contradicts continuity!
(By the intermediate-value theorem)

This example can be embedded into higher dimensions

Recourse sensitivity

Example



Impossibility result

General case

Theorem

If u_f is of the form $u_f(x, y) = \widetilde{u}(f(x), f(y))$ and if there exist $z_1, z_2 \in \mathbb{R}^d$ such that $\widetilde{u}(z_1, z_2) \geq \tau$ and $\widetilde{u}(z_1, z_1) < \tau$.

Then, there exists a continuous $f: \mathcal{X} \to \mathbb{R}$ for which no attribution method φ_f can be both recourse sensitive and robust.

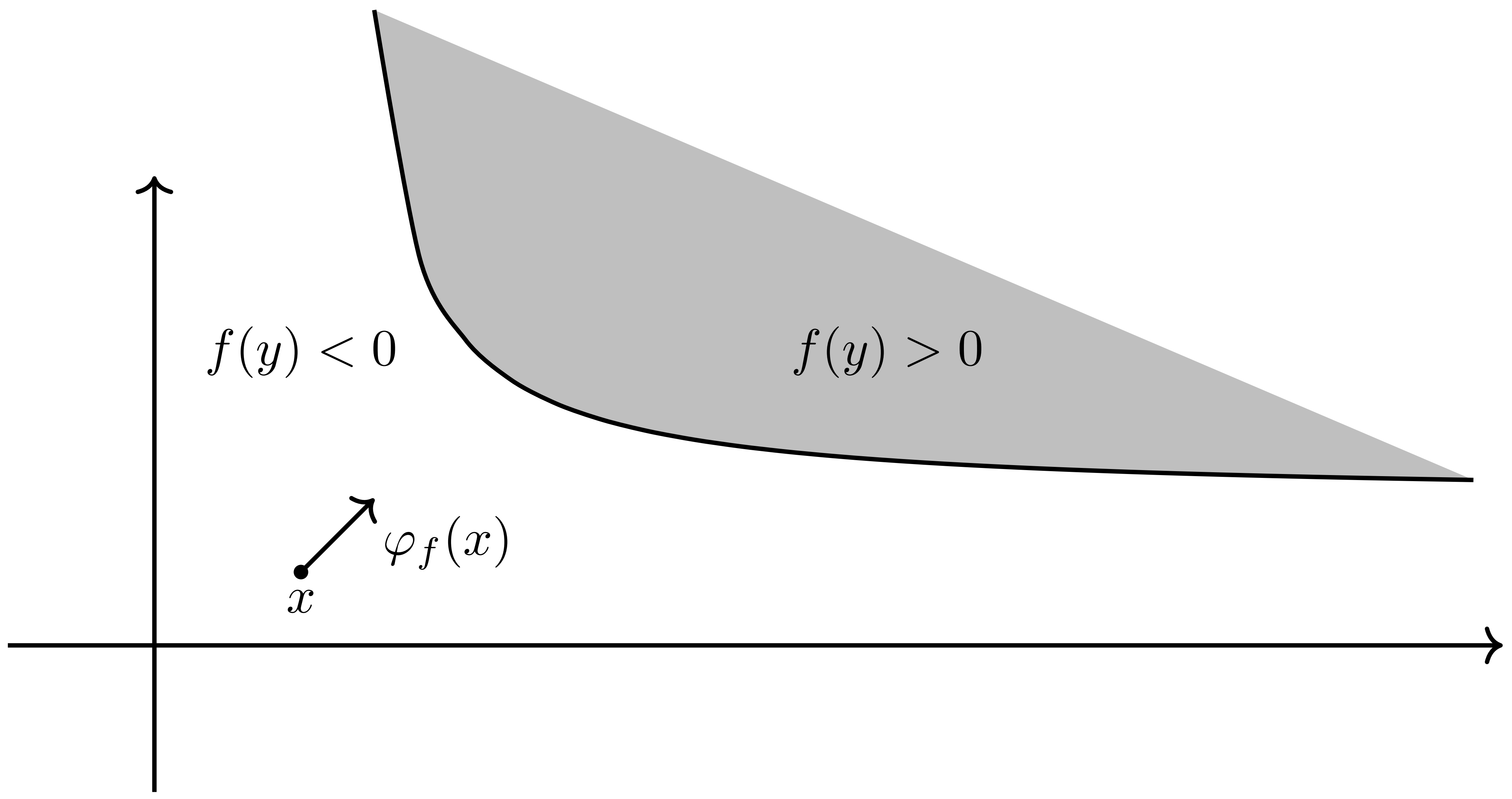
Attribution methods cannot always

- Provide recourse
- Be continuous

When is Recourse and Robustness possible?

Recourse and Robustness is possible sometimes
Binary classification

- Preferred class ($u_f(x, y) = f(y) \geq 0$)
- Let $U = \{x \mid f(x) > 0\}$ be convex
- Then Recourse and Robustness is possible!

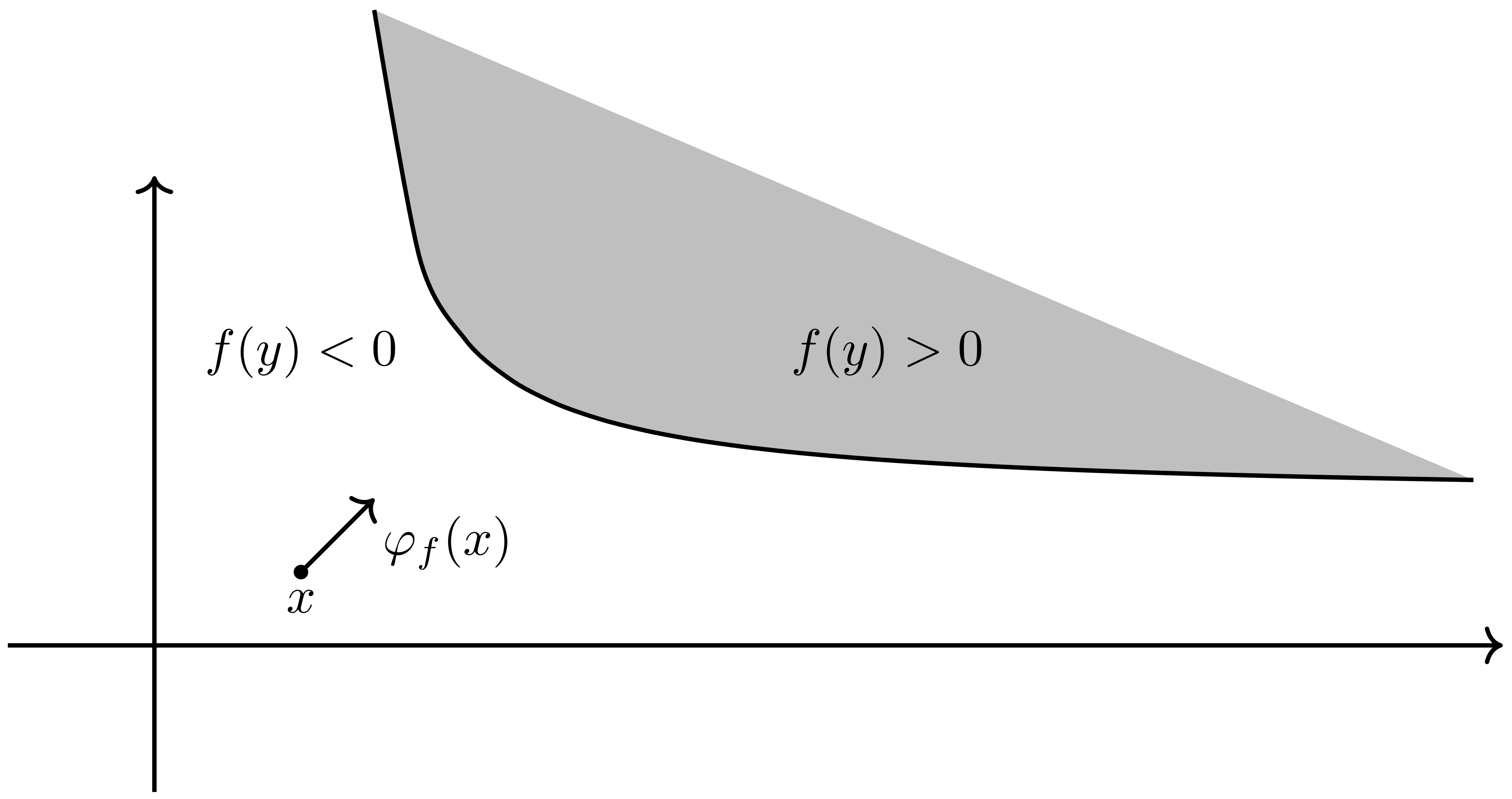


$$\varphi_f(x) = P_U(x) - x$$

Recourse and Robustness is possible sometimes

General case

- General Utility $u_f(x, y)$
- $U(x) = \{y \,|\, u_f(x, y) \geq \tau\}$ become x dependent
- We need:
  - "Continuity of $U(x)$"
  - Projections should exist and be unique



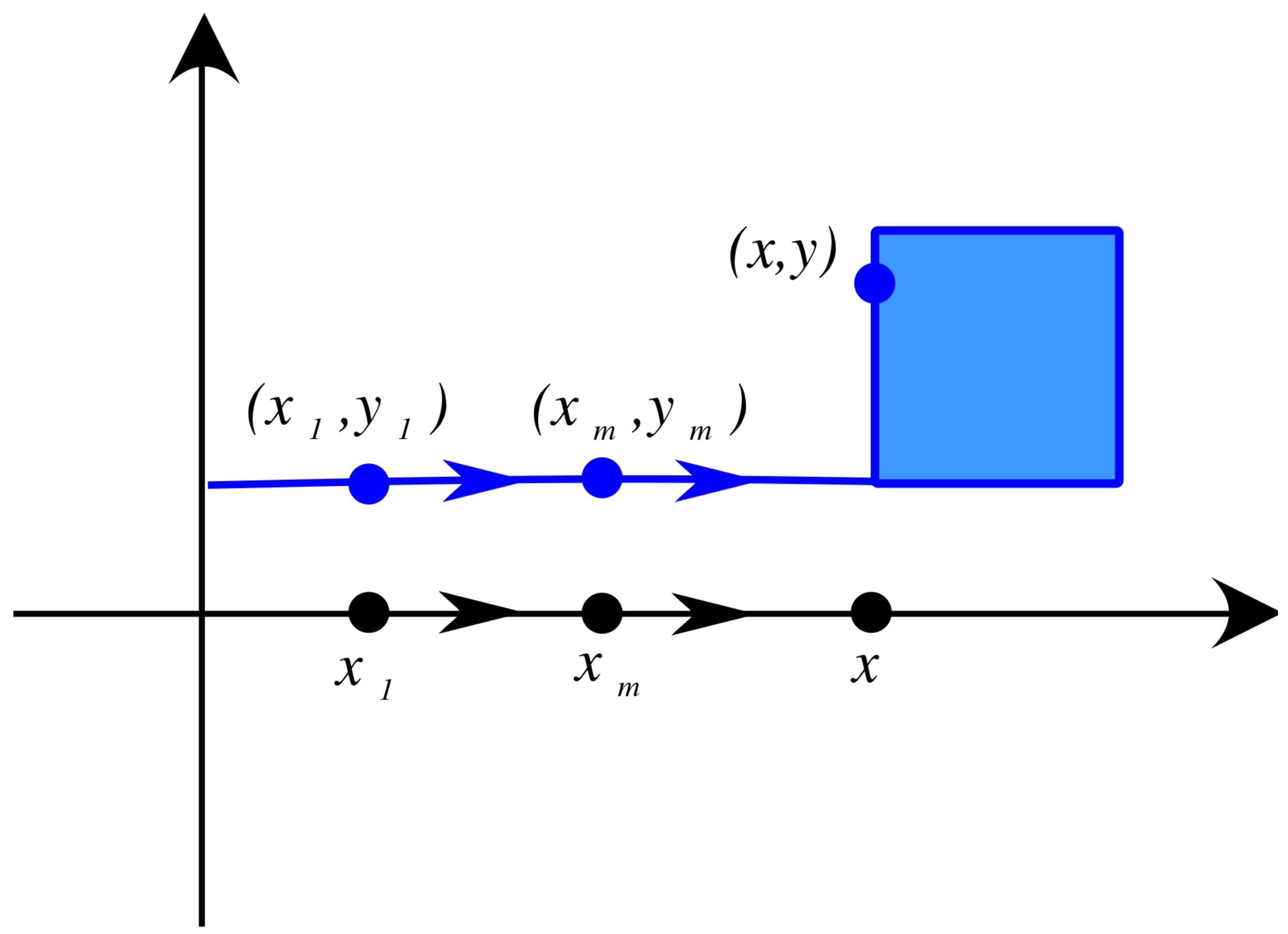
$$\varphi_f(x) = P_U(x) - x$$

Hemi-continuity

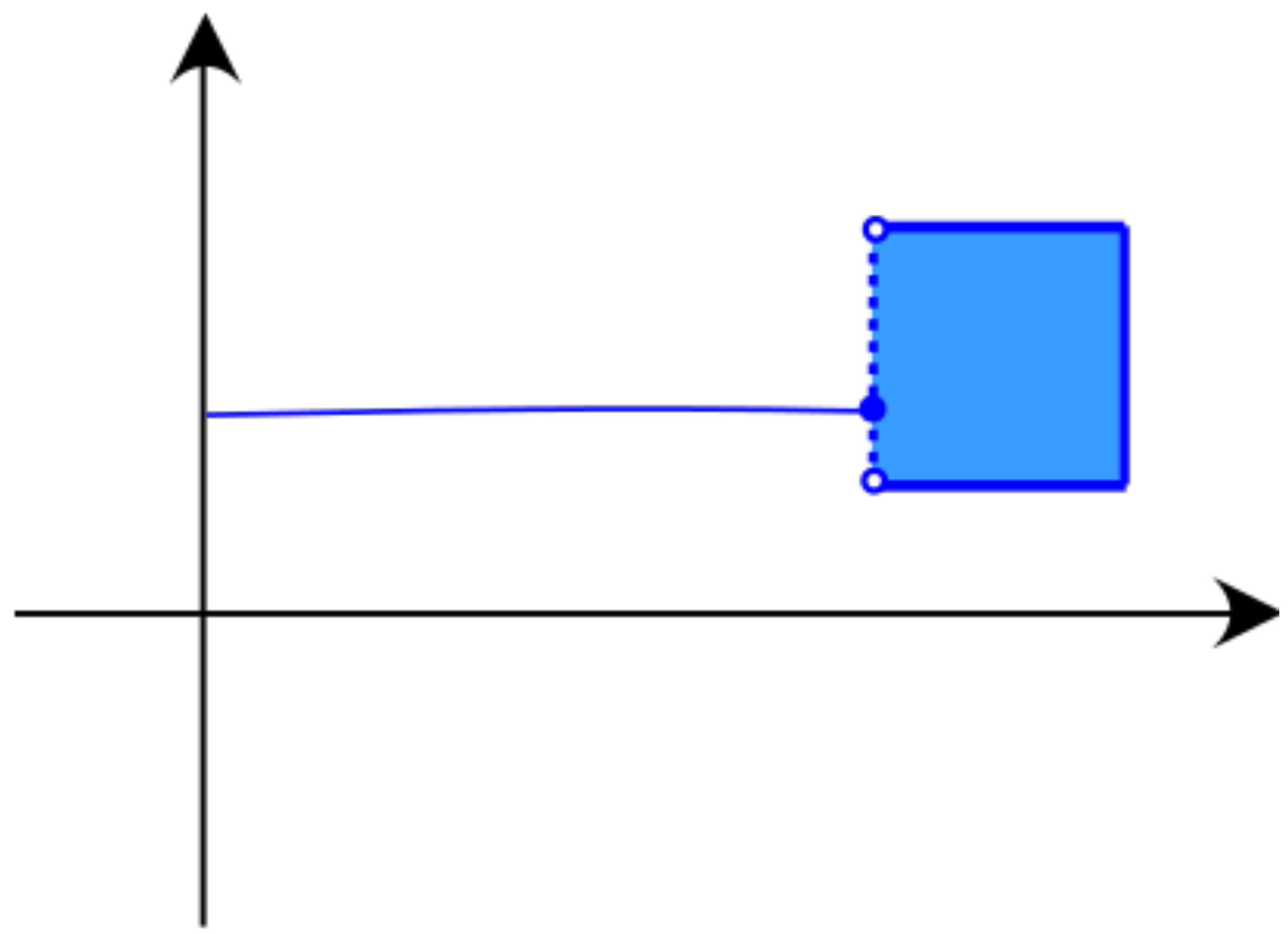
Set-valued function $U\colon \mathcal{X} \to 2^{\mathcal{Y}}$:

- Upper Hemi-continuity: U(x) cannot suddenly explode
- Lower Hemi-continuity: U(x) cannot suddenly implode

UHC, but not LHC[4]



LHC, but not UHC[4]



[5]Image source: https://en.wikipedia.org/wiki/Hemicontinuity

Conclusion

Summary:

- There exist f for which recourse sensitivity + robustness is impossible, for several machine learning tasks
- There are cases for which it is possible, but they require strong conditions
- Further extensions in the paper:
    - Sufficient Conditions for when Recourse and Robustness is possible
    - Full Characterisation for Single-Feature case
    - Discussion on possible ways around our impossibility result
    - Constraints on user actions

Thank you for your attention!

References

- Fokkema, Hidde, Rianne de Heide, and Tim van Erven. "Attribution-based Explanations that Provide Recourse Cannot be Robust." arXiv preprint arXiv:2205.15834 (2022).
- M. T. Ribeiro, S. Singh, and C. Guestrin. ”Why should I trust you?” explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.
- D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. ArXiv:1706.03825, 2017.
- Verma, Sahil, John Dickerson, and Keegan Hines. "Counterfactual explanations for machine learning: A review." arXiv preprint arXiv:2010.10596 (2020).

Recourse and Robustness is possible sometimes

General case

Theorem

Let $\delta > 0$, $\tau \geq 0$, $f\colon \mathcal{X} \to \mathbb{R}$ be a continuous function and $u_f(x, y)$ a utility function with the following properties:

1. For every $x \in \mathcal{X}$, the projection onto $U(x)$ exists and is unique;
2. The set-valued function $U(x)$ is Hemi-continuous and closed.

Then the function given by:

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

Is a recourse sensitive and robust attribution map.

Proof idea:

- Berge's Maximum Theorem gives Continuity almost immediately
- Check that Recourse sensitivity is satisfied

Impossibility result

Proof sketch

Define

- Take z_1, z_2 such that $\widetilde{u}(z_1, z_2) \geq \tau$.
- $L = \{x \in \mathscr{X} \mid$ there exists some $y \in [x - \delta, x]$ with $u_f(x, y) \geq \tau\}$,
- $R = \{x \in \mathscr{X} \mid$ there exists some $y \in [x + \delta, x]$ with $u_f(x, y) \geq \tau\}$.