Attribution-based Explanations that Provide Recourse Cannot be Robust

Hidde Fokkema (University of Amsterdam)

Seminar on the Theory of Interpretability

2024-05-07

Joint Work

- All work presented was created in collaboration with:

Dr. Tim van Erven
University of Amsterdam

Dr. Rianne de Heide
Vrije Universiteit

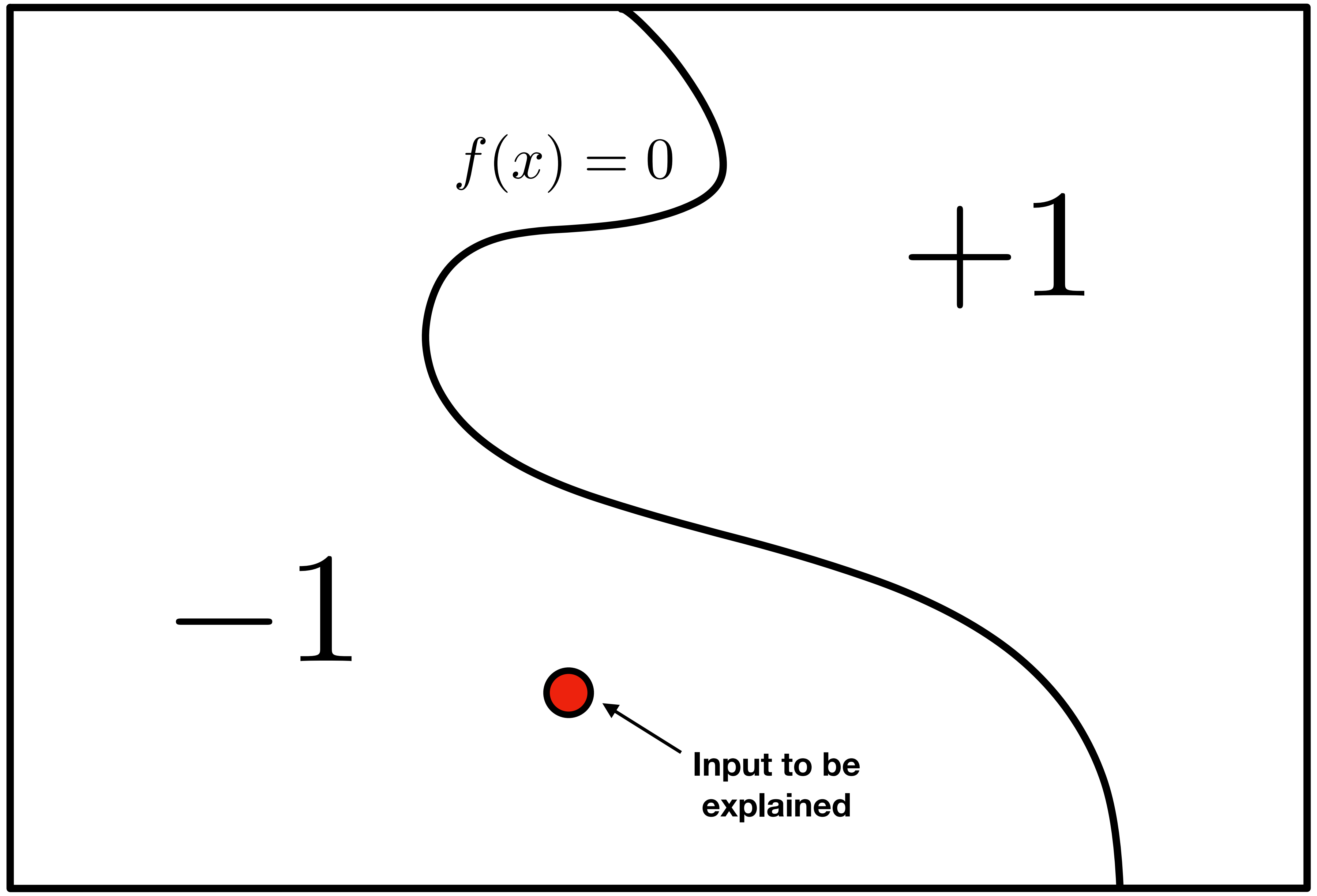
Outline

- Attribution & Counterfactual 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:

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

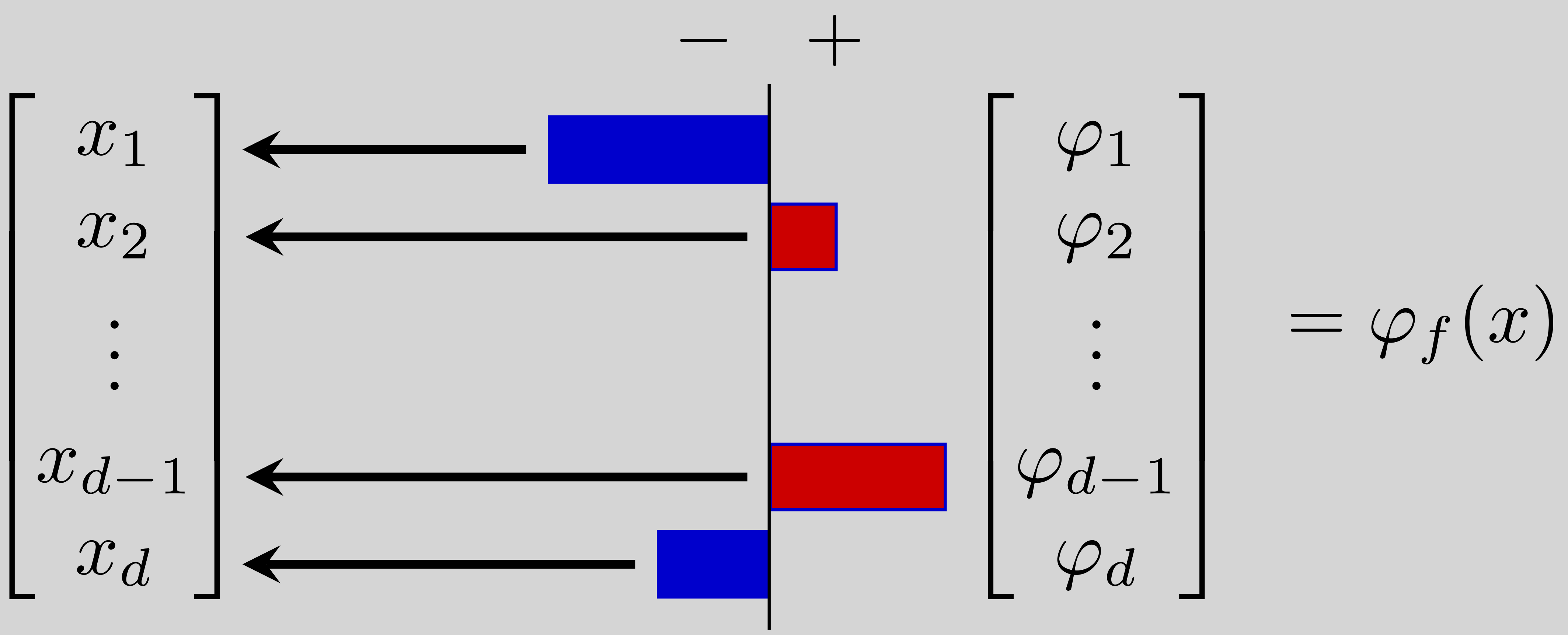
- **Local:** Only explain the part of f that is relevant for x
- **Post-Hoc:** The function f is given and fixed

Setting

Attribution methods

Machine learning model, e.g. a classifier:

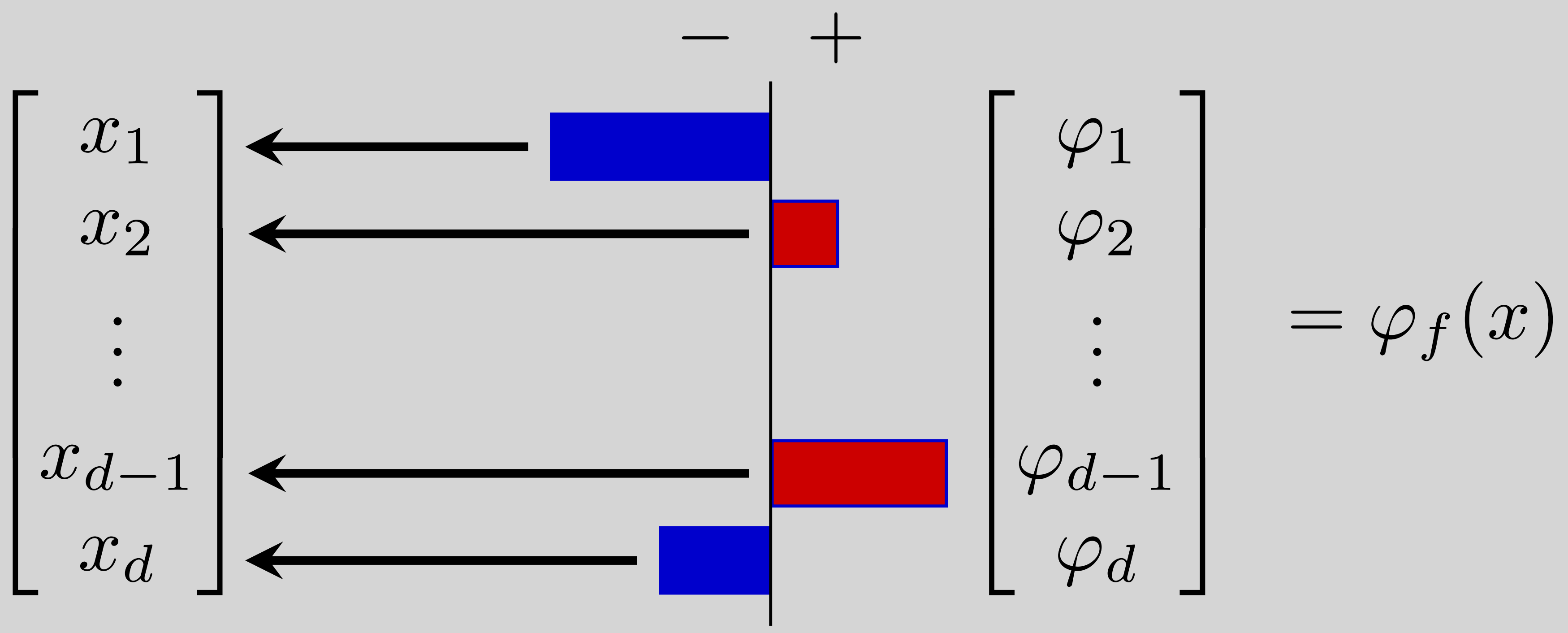
$$f : \mathcal{X} \subseteq \mathbb{R}^d \to [0,1], \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \mapsto y$$



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

Example

Attribution methods



f linear, low dimension d

$$f(x) = \theta_0 + \sum_{i=1}^{d} x_i \theta_i$$

$$\varphi_f(x)_i = \theta_i$$

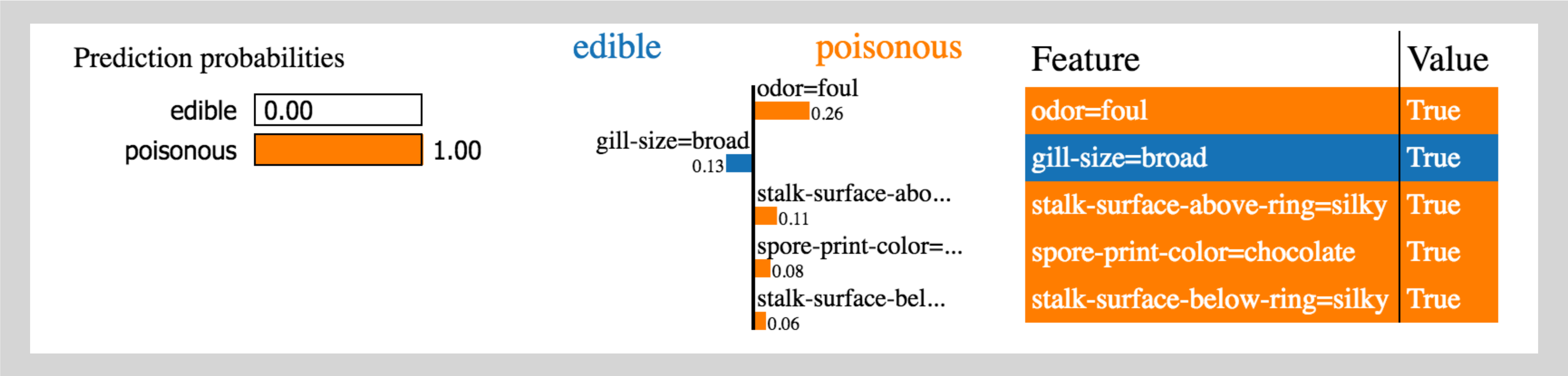
- In general φ_f will depend on x
- Most methods follow this example, by locally linearising around x

Examples (LIME)

<u>Lime</u>: Extract local linear approximations of f near x and report coefficients

- Optionally: Apply some dimension reduction before linearising.

Lime for tabular data[1]



[1] Image source: https://github.com/marcoctr/lime

Examples (Gradient)

Gradient methods

- Vanilla gradient:

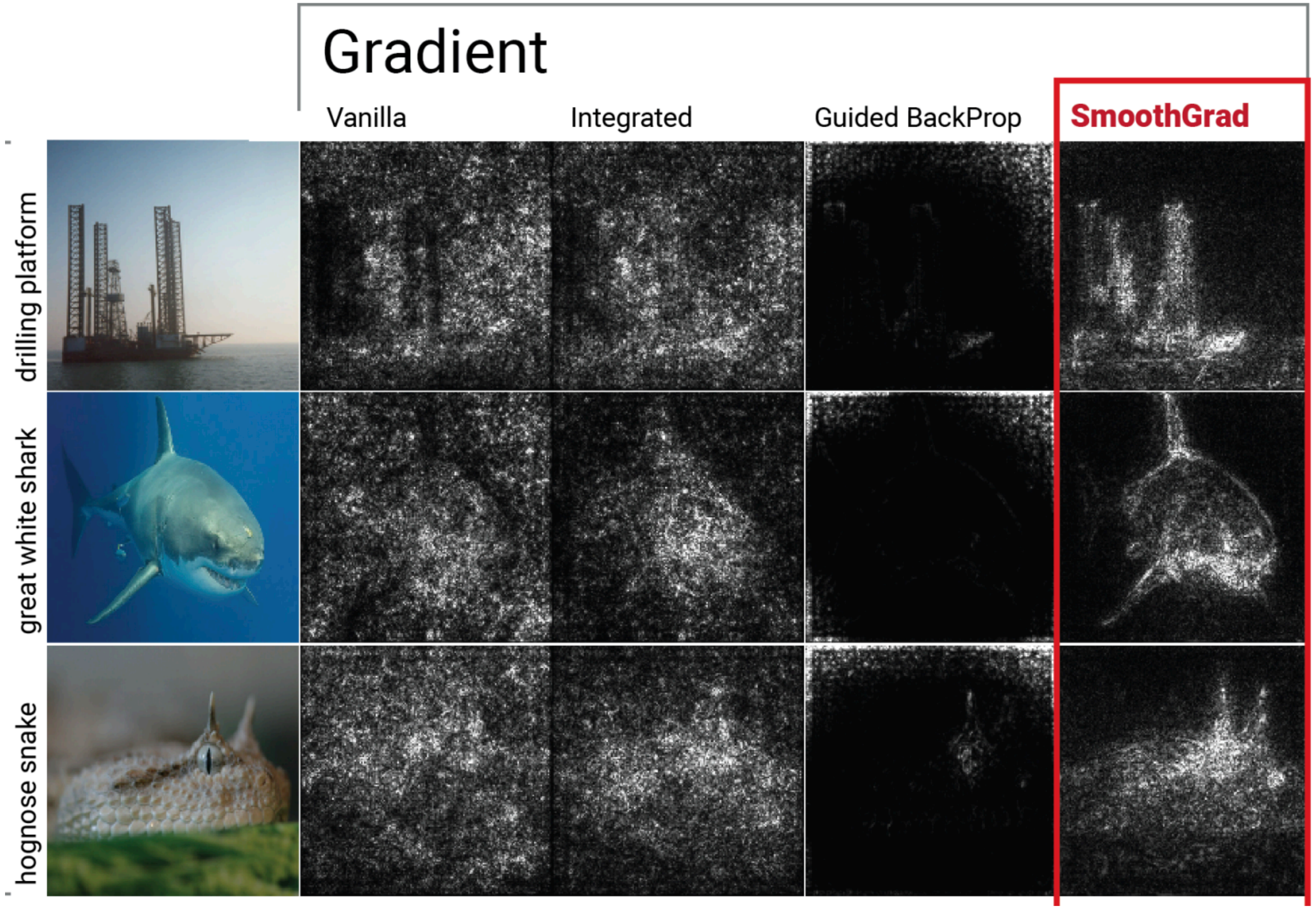
$$\varphi_f(x) = \nabla f(x)$$

- SmoothGrad:

$$\varphi_f(x) = \mathbb{E}_{Z \sim \mathcal{N}(x,\Sigma)}[\nabla f(Z)]$$

- Integrated Gradients:

$$\varphi_f(x) = (x - x_0)\int_0^1 \nabla f(x_0 + t(x - x_0))\mathrm{d}t$$

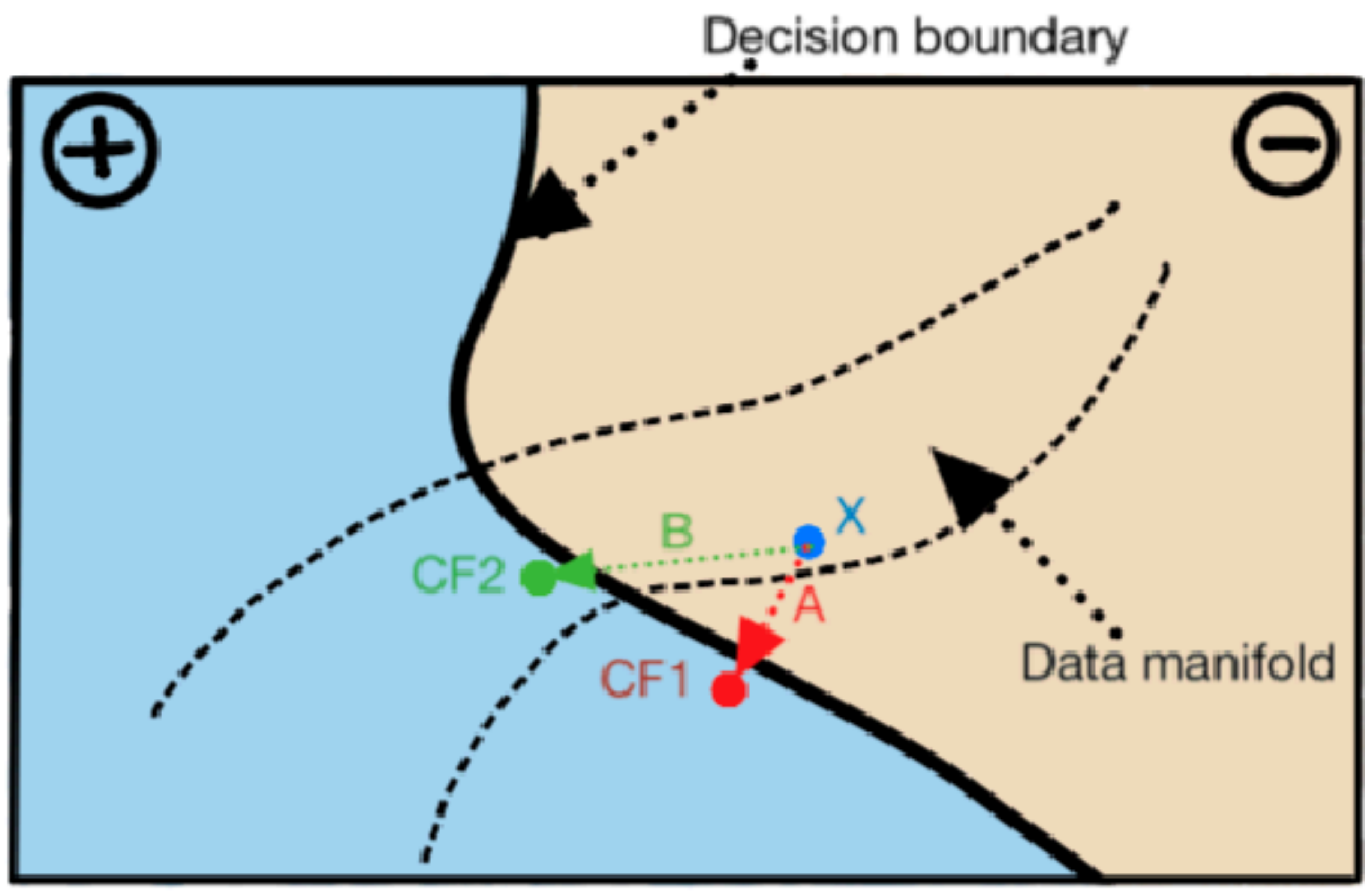


[4] Image source: [Smilkov et al., 2017]

Counterfactuals

Example

"If your Income would have been €40.000,- instead of €35.000,-, your loan application would have been accepted”



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

Counterfactuals as attributions

Counterfactuals

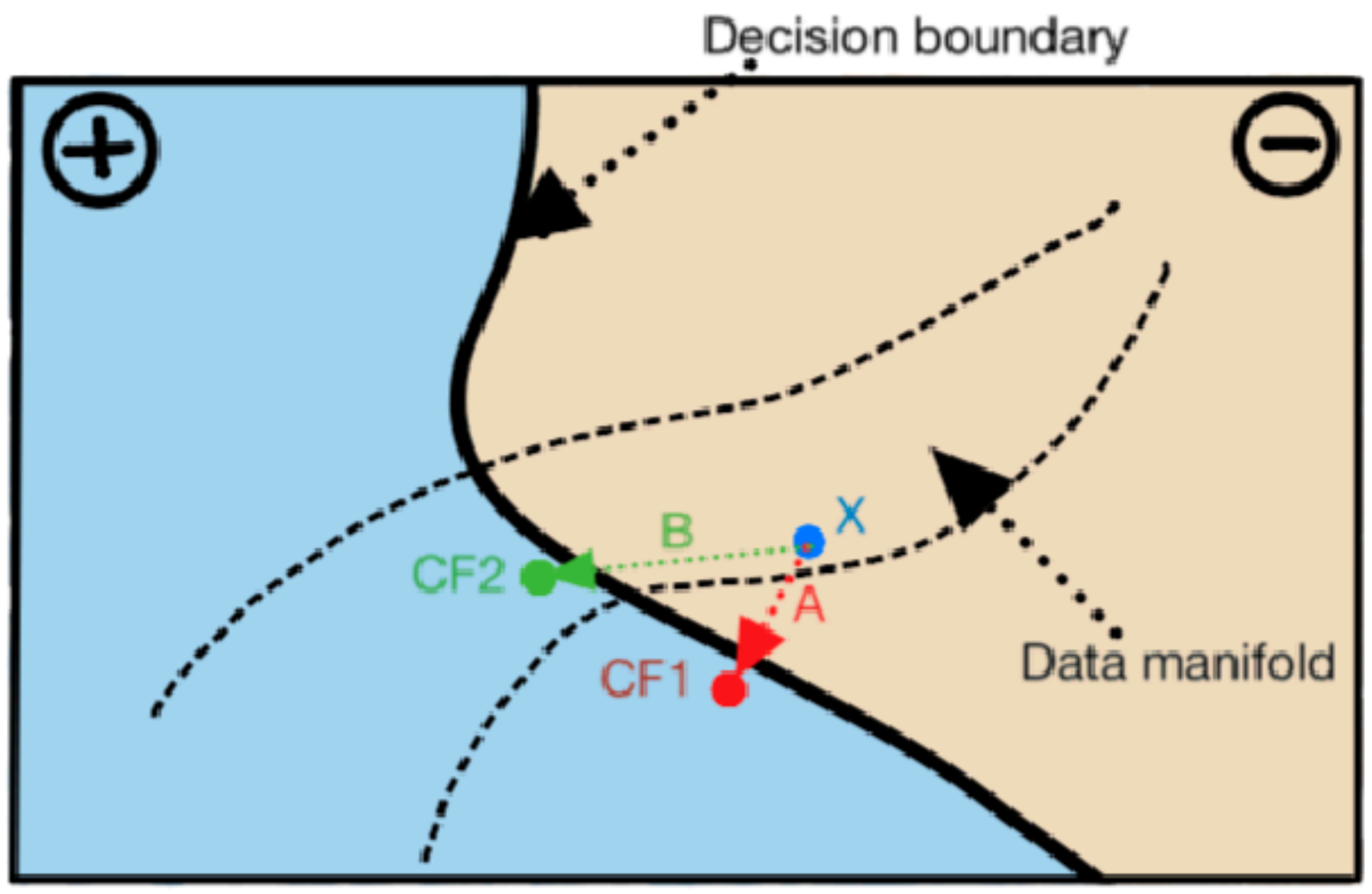
Consider Binary classification $f\colon \mathcal{X} \to \{-1,1\}$ and let $x \in \mathcal{X}$.

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

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



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

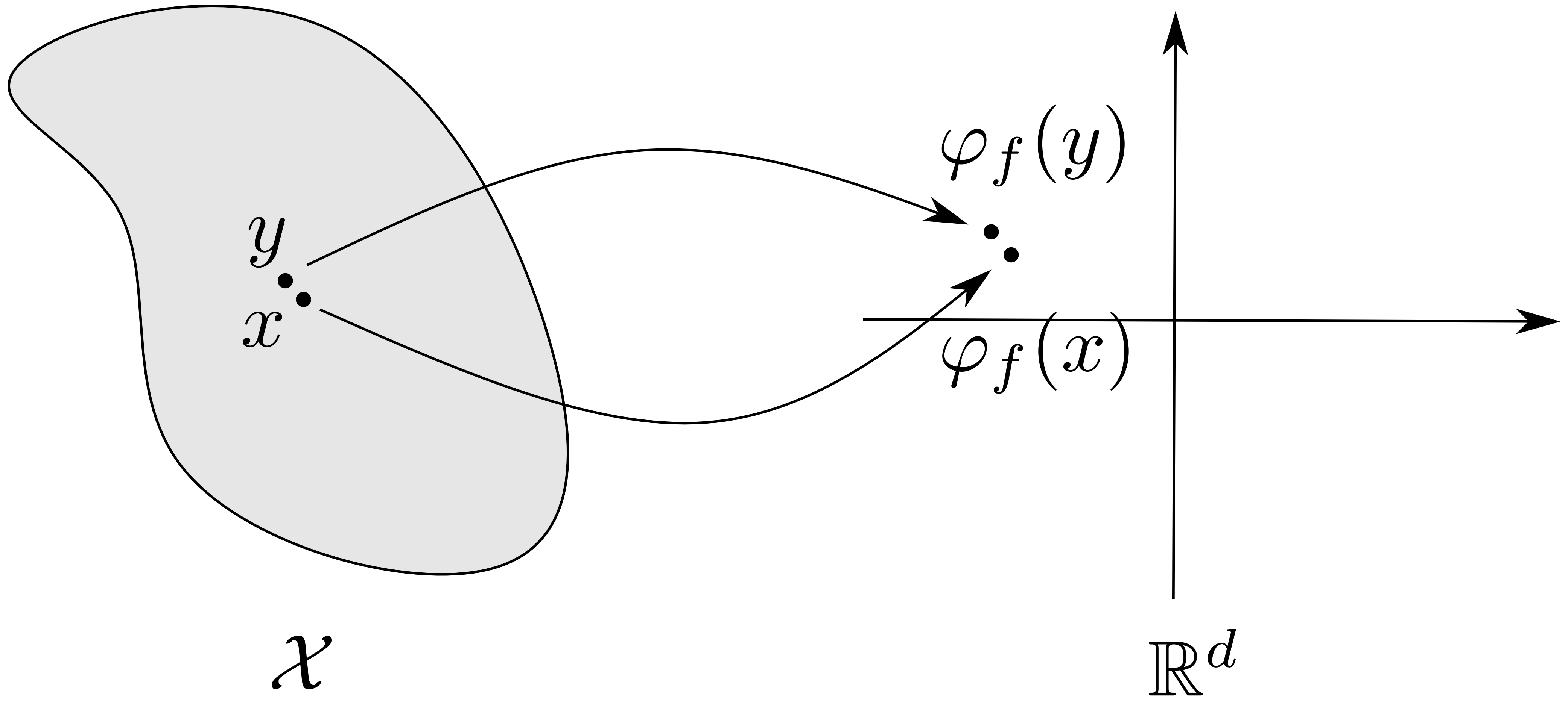
Robustness &
Recourse sensitivity

Robustness

Similar users require similar explanations

Definition

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



Similar definitions and motivation can be found in [Karimi et al 2021, Alvarez-Melis and Jaakkola 2018, Khan et al. 2024]

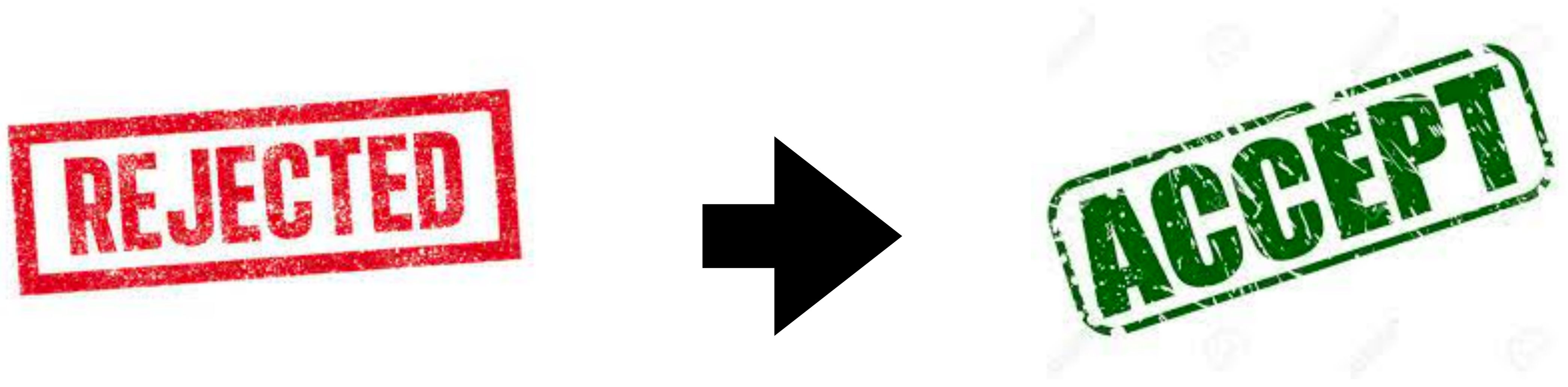
Recourse sensitivity

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

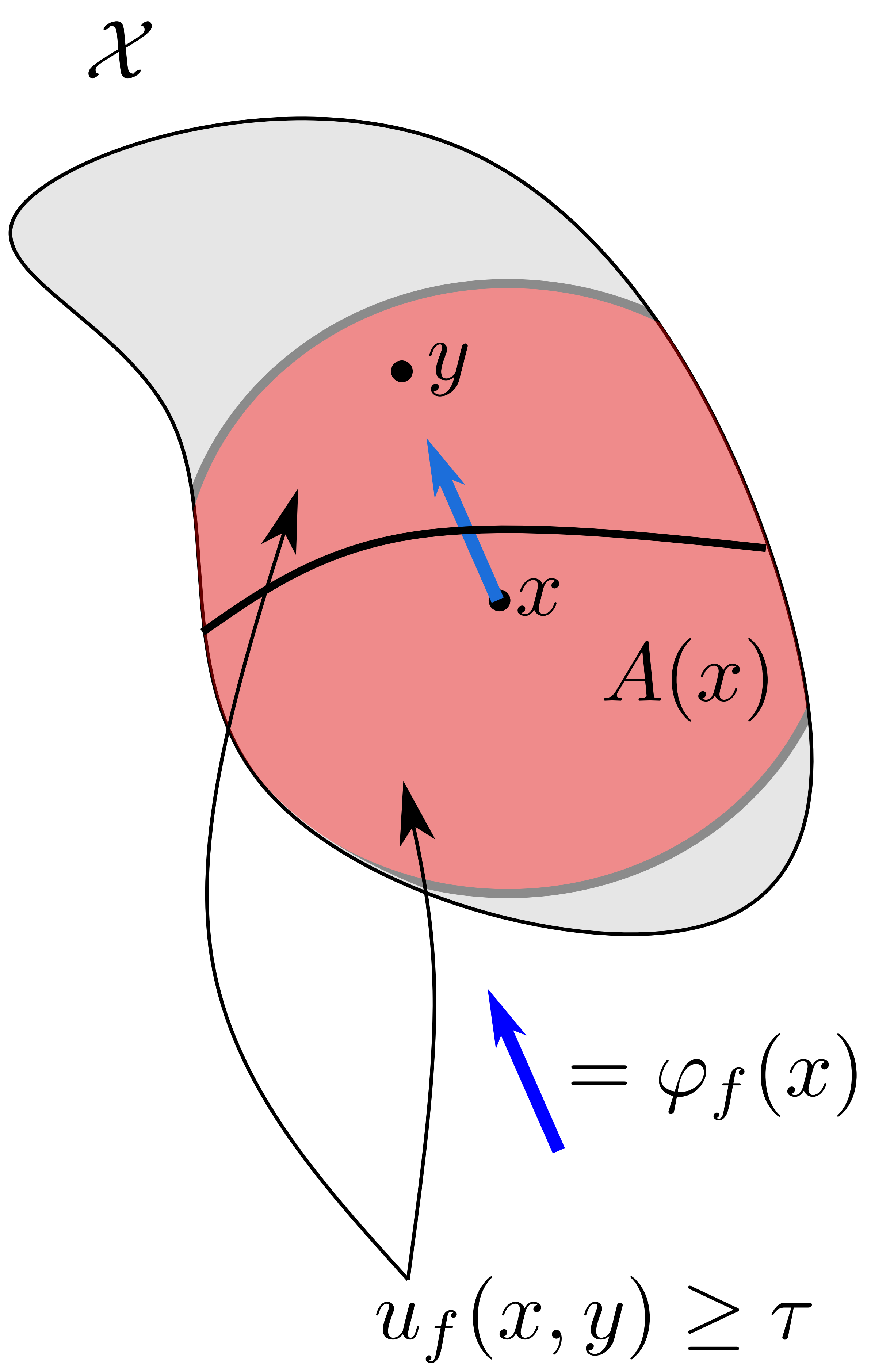


Recourse sensitivity

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 requirement!



Recourse sensitivity

Utility

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

- Undesirable classification:

$$u_f(x, y) = f(y) \geq 0$$

- Increase score:

$$u_f(x, y) = f(y) - f(x) \geq \tau$$

- Decrease a probability:

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

Recourse sensitivity

Definition

Define set of attainable points from x

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

Definition

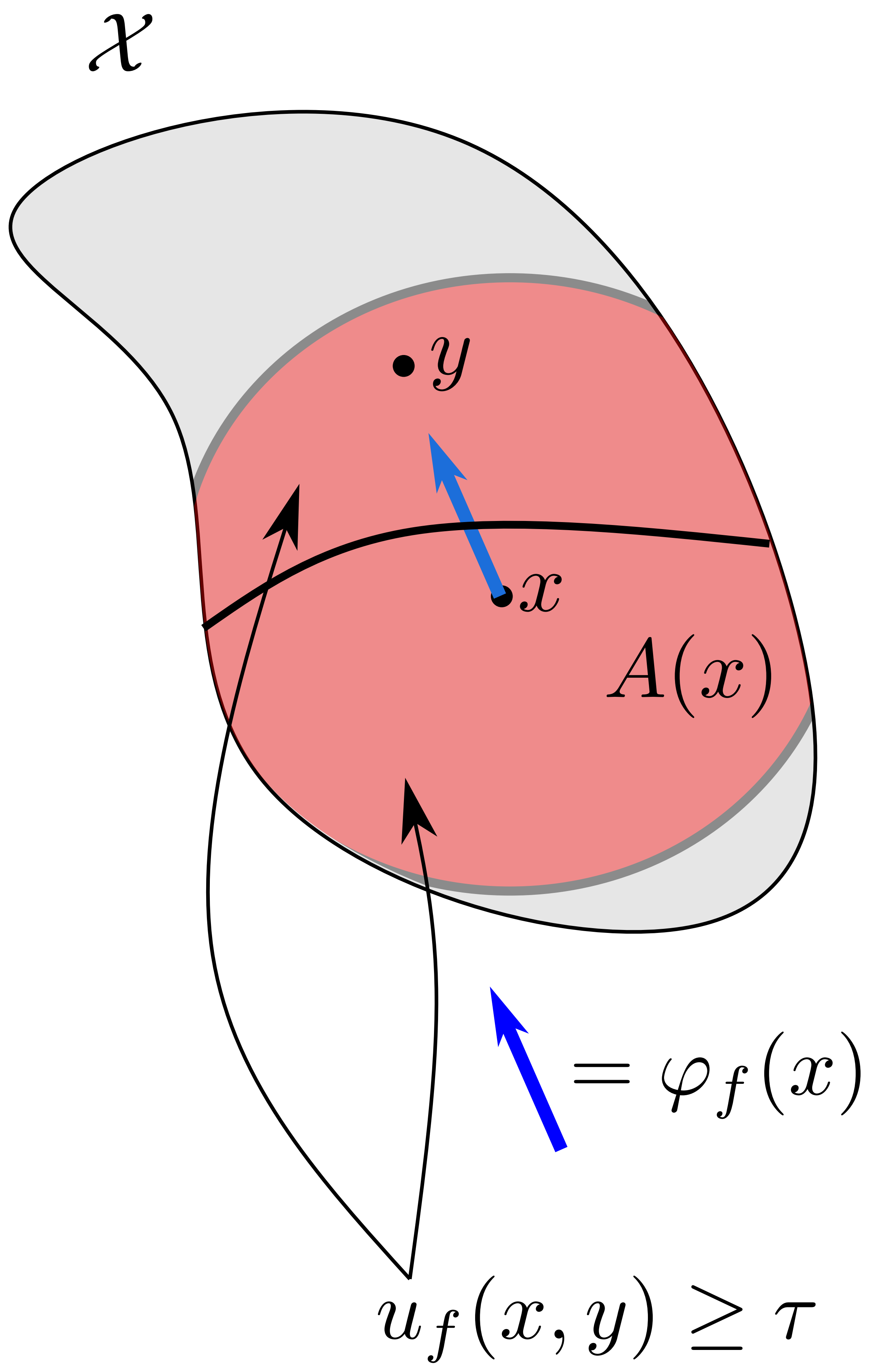
Consider the points close to x that achieve sufficient utility

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

An Attribution function φ_f is called ***Recourse Sensitive*** if for each $x \in \mathcal{X}$ there exists $\alpha > 0$ and $y \in U(x)$ s.t.

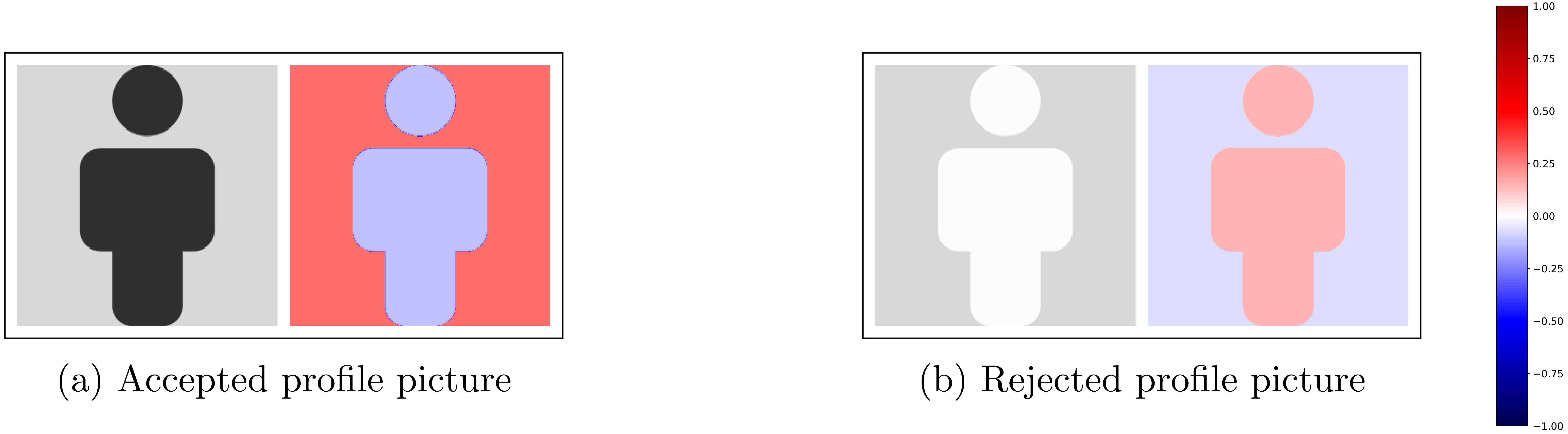
$$\varphi_f(x) = \alpha(y - x),$$

If $U(x) \neq \emptyset$.



Recourse sensitivity

Example



(a) Accepted profile picture

(b) Rejected profile picture

Impossibility

Impossibility result

Attribution methods cannot always

- Provide Recourse
- Be Robust

Impossibility result

Specific case (Binary classification)

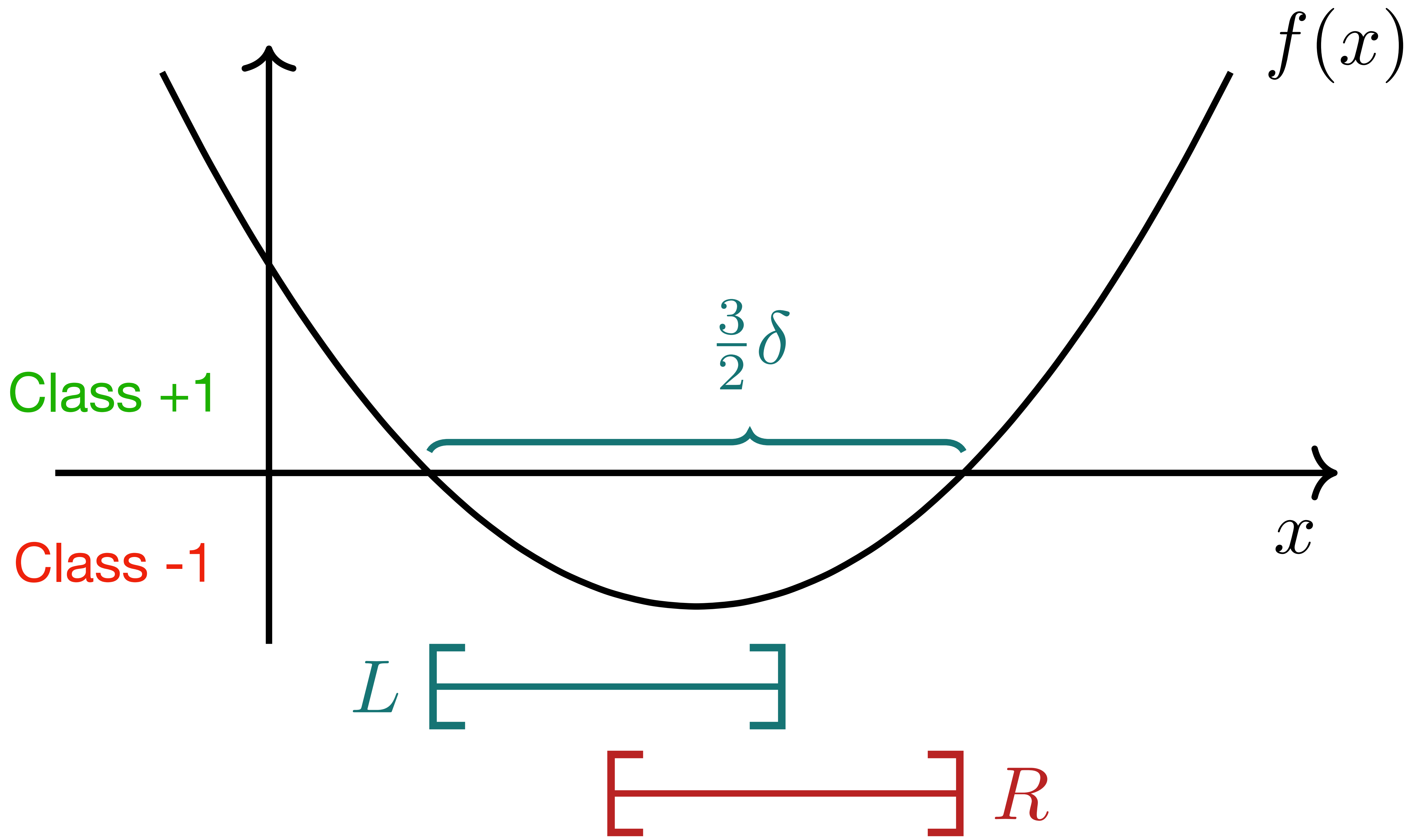
Setting

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

Theorem

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

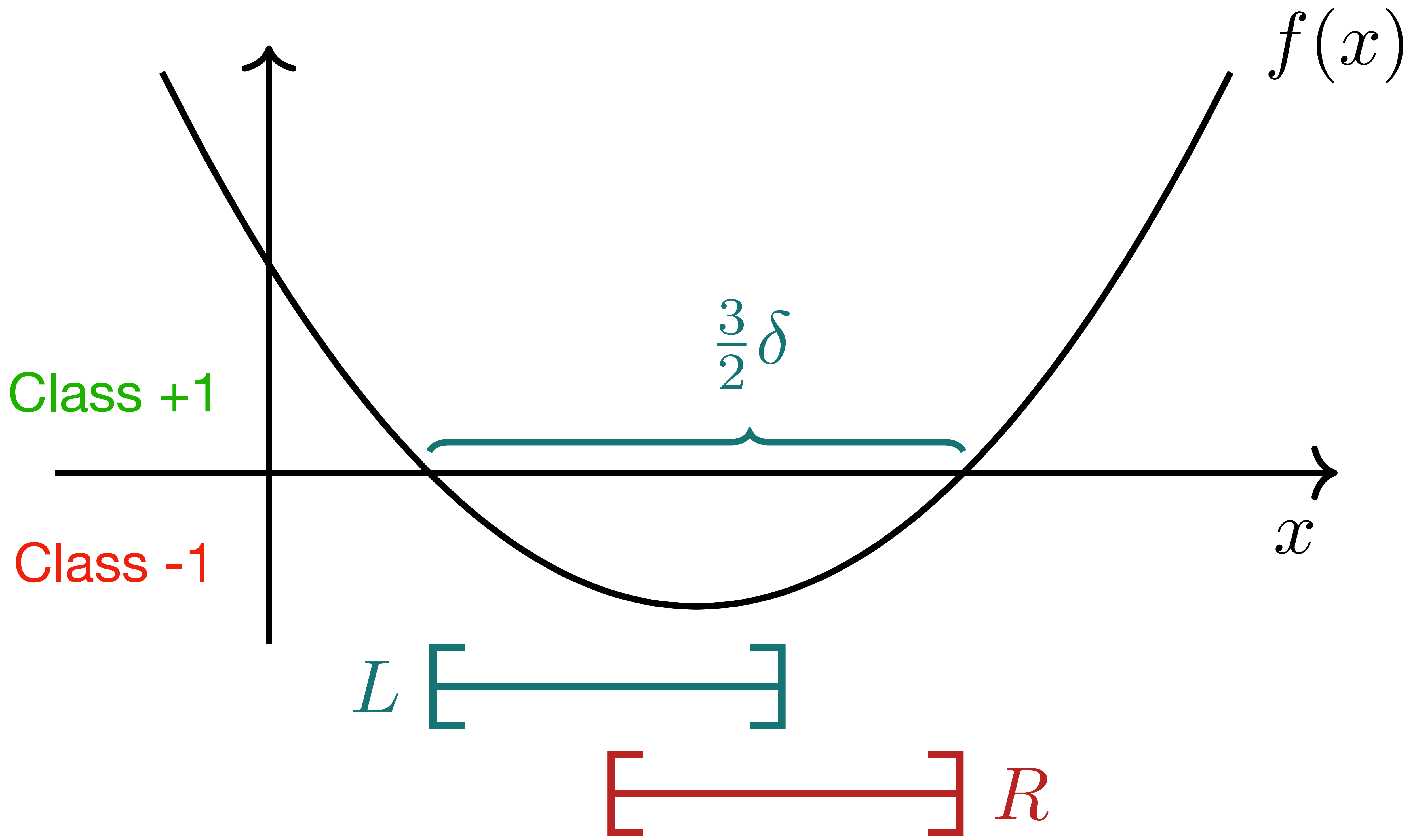
Proof sketch



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

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

Proof sketch

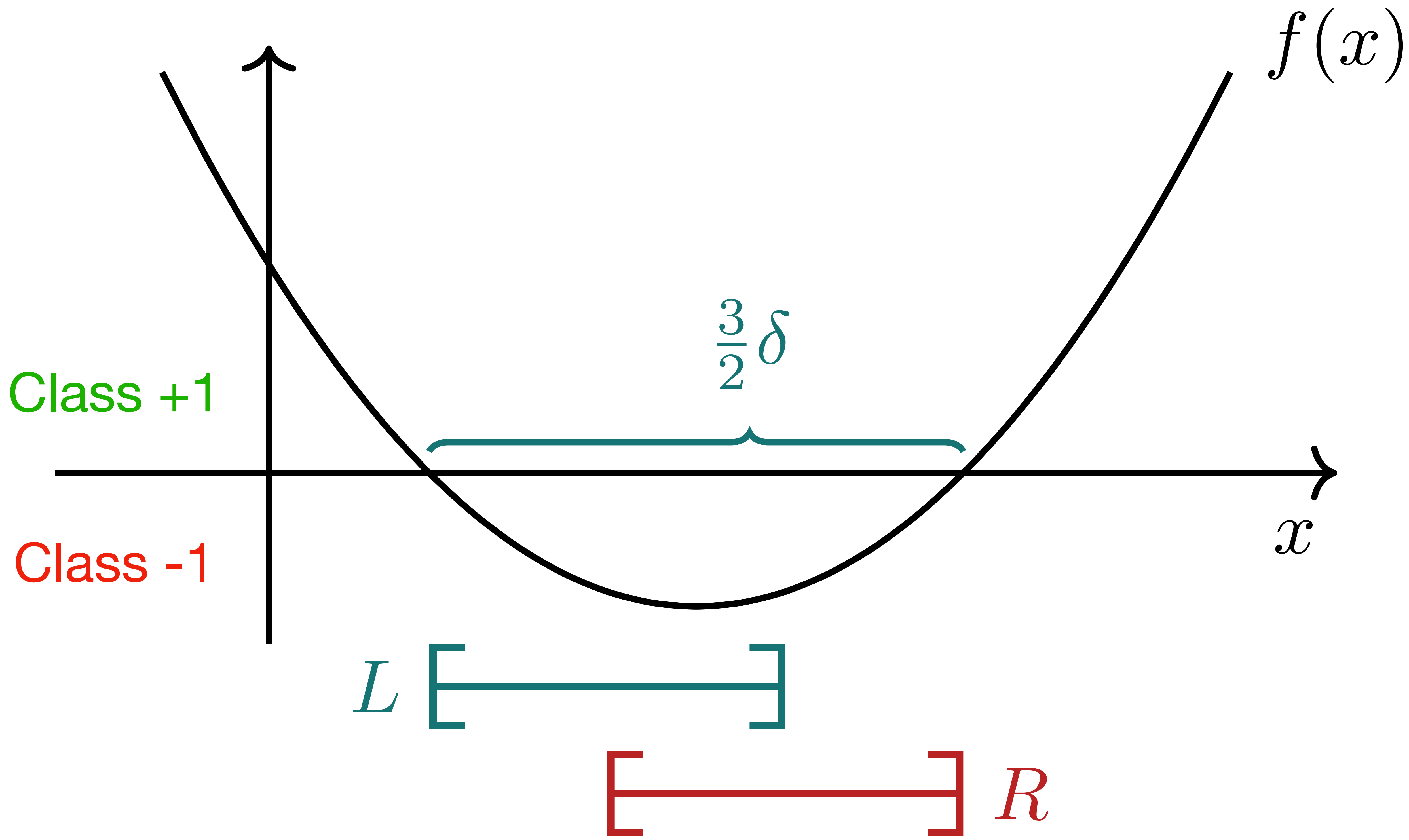


$R = \{x \mid \text{recourse is possible by moving at most } \delta \text{ left}\}$

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

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

Proof sketch



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

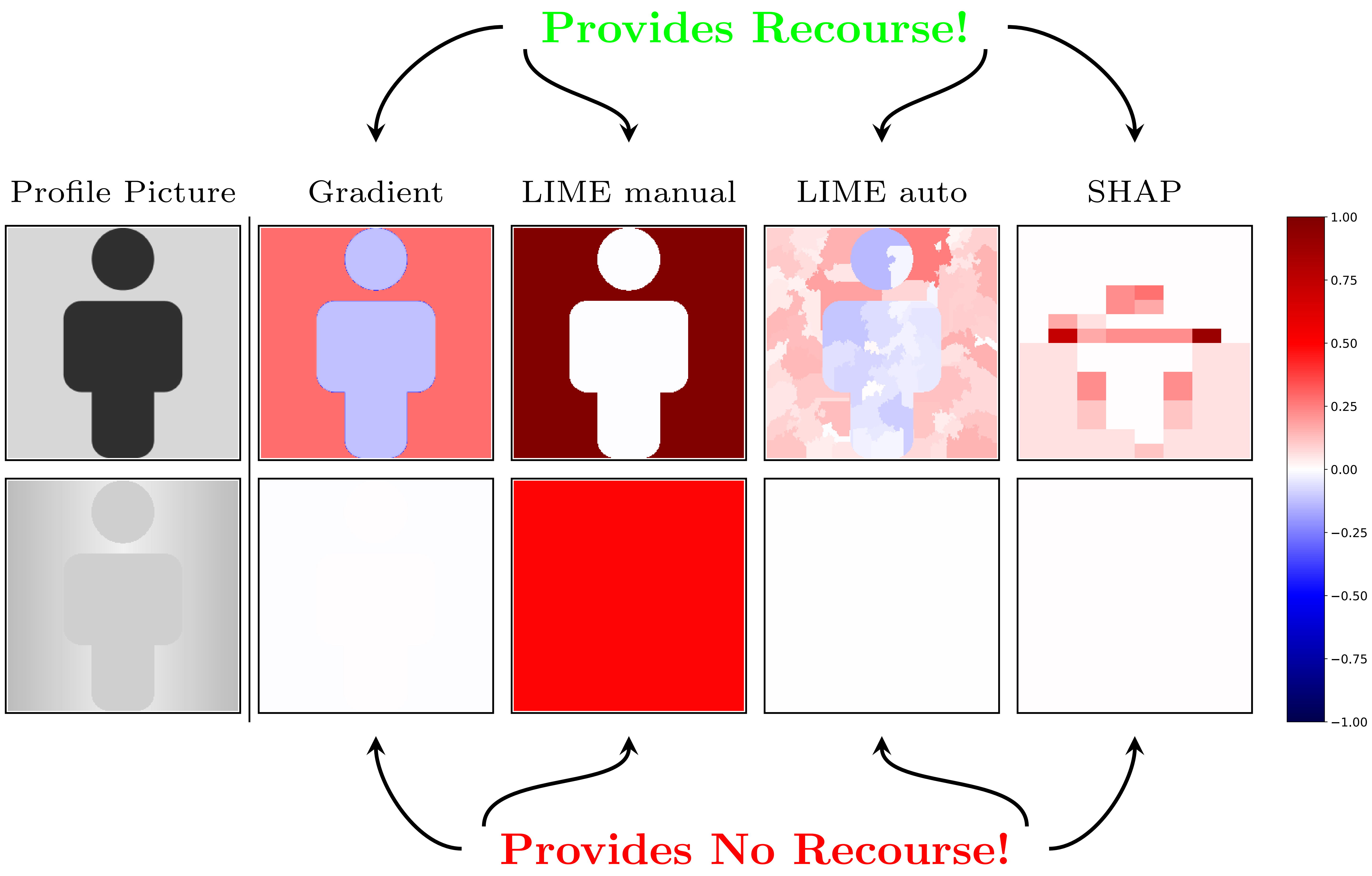
$$\varphi_f(x) = \begin{cases} < 0 & \text{for } x \in L \backslash R \\ > 0 & \text{for } x \in R \backslash L \\ \neq 0 & \text{for } x \in L \cap R \end{cases}$$

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

This example can be embedded into higher dimensions

Recourse sensitivity

Example

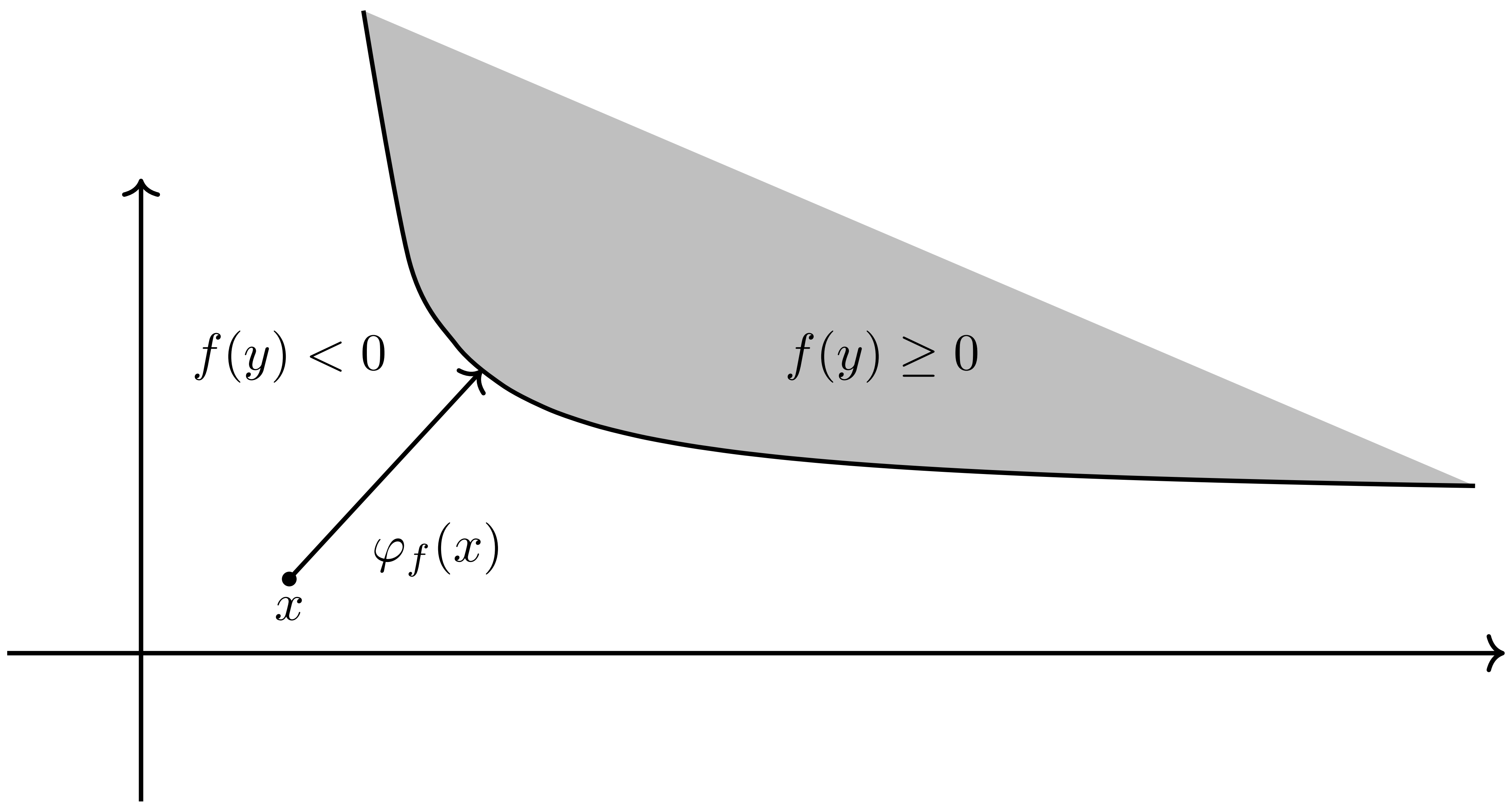


When is Recourse and Robustness possible?

Recourse and Robustness is possible sometimes

Binary classification

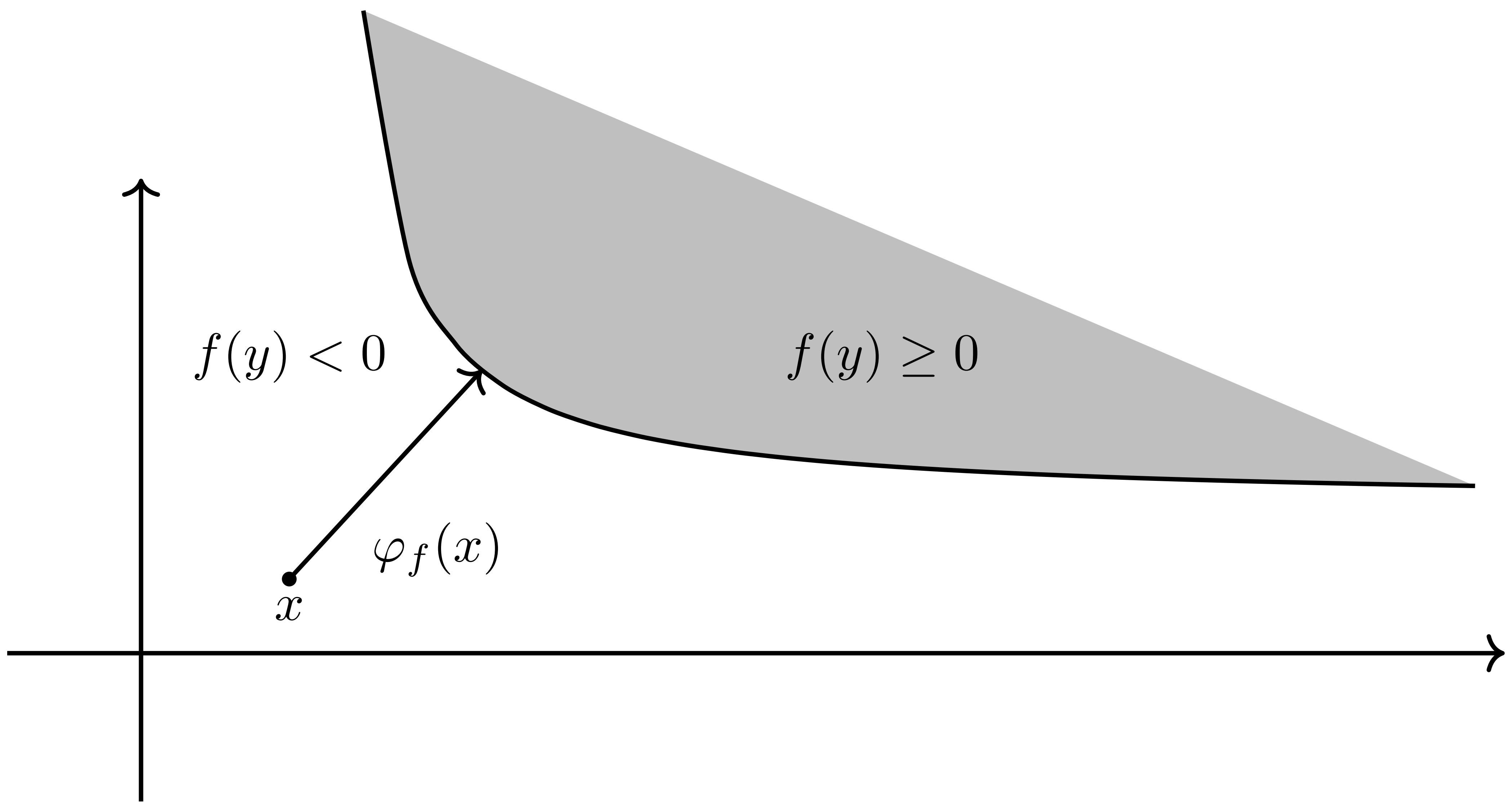
- Preferred class, i.e. $u_f(x, y) = f(y) \geq 0$
- Let $U = \{x \in \mathcal{X} \mid f(x) \geq 0\}$ be convex
- Then Recourse and Robustness is possible!



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

Recourse and Robustness is possible sometimes
General case

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



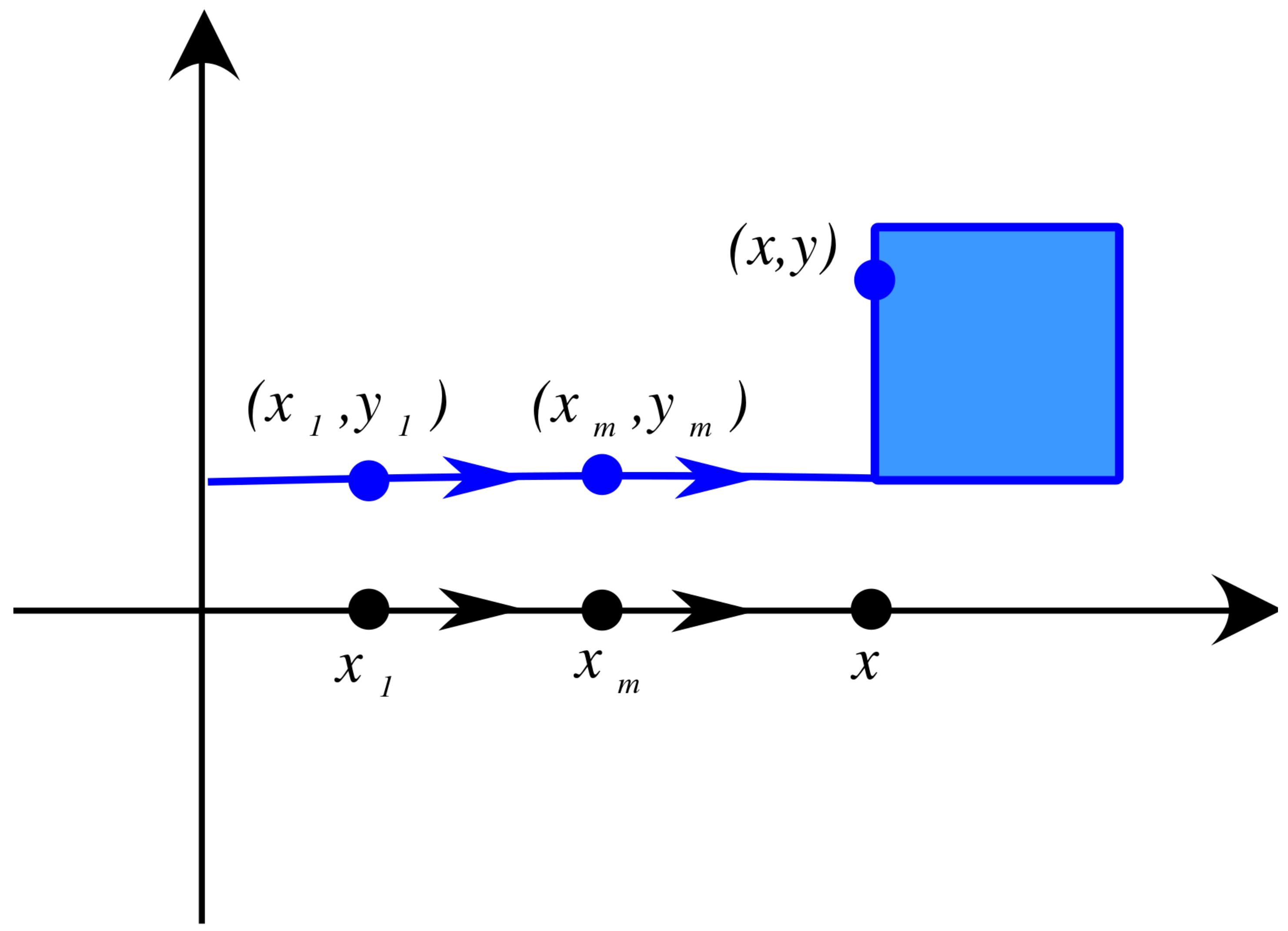
$$\varphi_f(x) = P_U(x) - x$$

Hemi-continuity

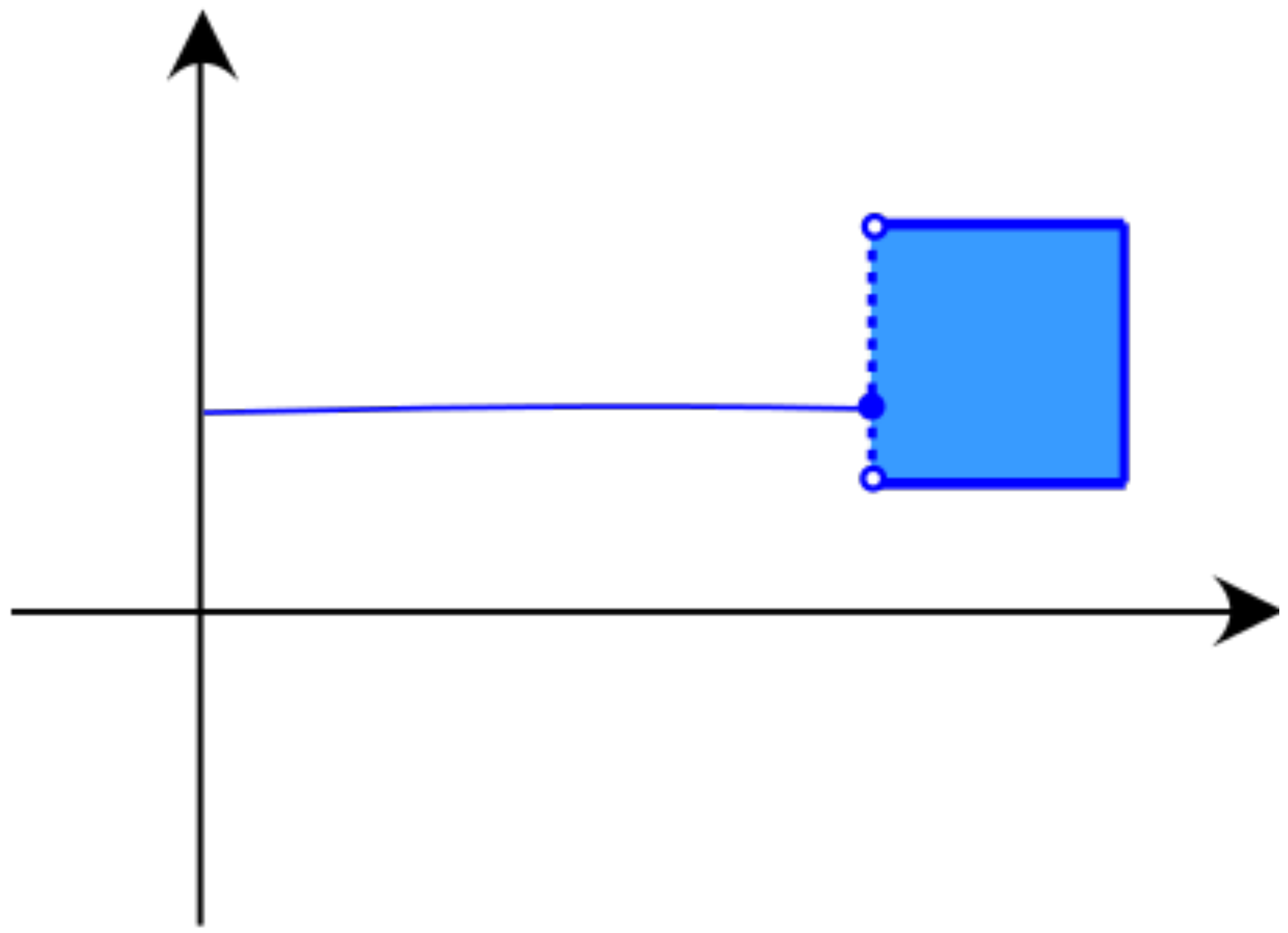
Set-valued function $U\colon \mathscr{X} \to 2^{\mathscr{Y}}$:

- 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]



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

Recourse and Robustness is possible sometimes

General case

Theorem

Let $u_f(x, y)$ be 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 sketch:

- Berge's Maximum Theorem gives continuity of the projections
- Check that Recourse sensitivity is satisfied

Work-arounds

What if we change the set up a bit?

Some observations:

- Counterfactuals are always recourse sensitive
- Robustness only fails if the counterfactual is not unique

Possible work-arounds

1. **Set-Valued explanations:** Give the user all possible ways to achieve the goal:
   - Pro: Recourse & Robustness is possible
   - Con: Computational problems & loses Attribution interpretation
2. **Linearising with High-level features/ Concepts:** Attribute groups of features/ concepts in the features:
   - Pro: Attribution, Recourse & Robustness is possible
   - Con: Definition of concepts ambiguous / combinatorial explosion of feature groups

Conclusion

Summary:

- There exist f for which recourse sensitivity + robustness is impossible
- There are cases for which it is possible, but they require strong conditions
- Sufficient Conditions for when Recourse and Robustness is possible
- Discussion on possible ways around this Impossibility result

Further extensions in the paper:

- Full characterisation in Single Feature case
- 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. JMLR vol. 24, pp 1-37 (2023).**

- D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods. In Proceedings of the 2018 Workshop on Human interpretability in Machine Learning. ICML, 2018.
- A-H Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Computing Surveys (CSUR), 2021.
- Khan, Z. Q., Hill, D., Masoomi, A., Bone, J. T., & Dy, J. (2024, April). Analyzing Explainer Robustness via Probabilistic Lipschitzness of Prediction Functions. In *International Conference on Artificial Intelligence and Statistics* (pp. 1378-1386). PMLR.
- 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).
- Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
- Leavitt, Matthew L., and Ari Morcos. "Towards falsifiable interpretability research." arXiv preprint arXiv:2010.12016 (2020).