Probabilistic Explanations for Regression Models

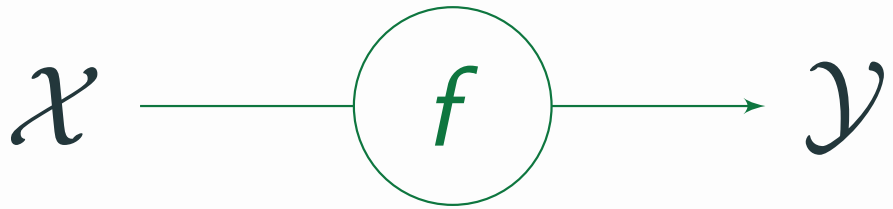
Frederic Koriche, Jean-Marie Lagniez, and **Chi Tran**

uai2025 — Rio de Janeiro, Brazil — 21-25 July, 2025

Univ. Artois, Centre de Recherche en Informatique de Lens (CRIL)
AI Chair EXPEKCTATION of the French Agency of National Research

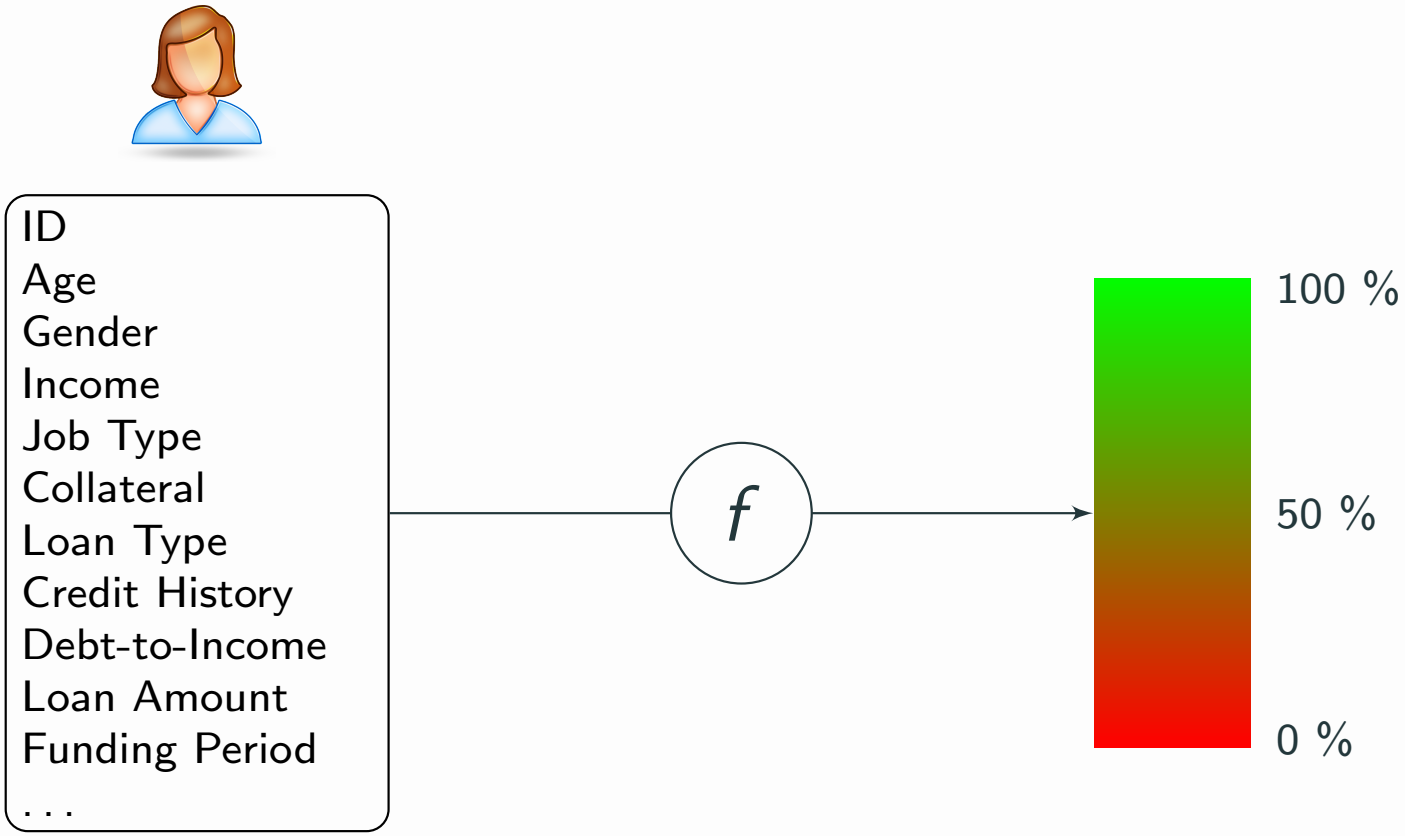
Outline

Regression Models



Conceptually, a regression model is a mapping f from a set $\mathcal{X} \subseteq \mathbb{R}^d$ of *data instances* to a set $\mathcal{Y} \subseteq \mathbb{R}$ of *outcomes*.

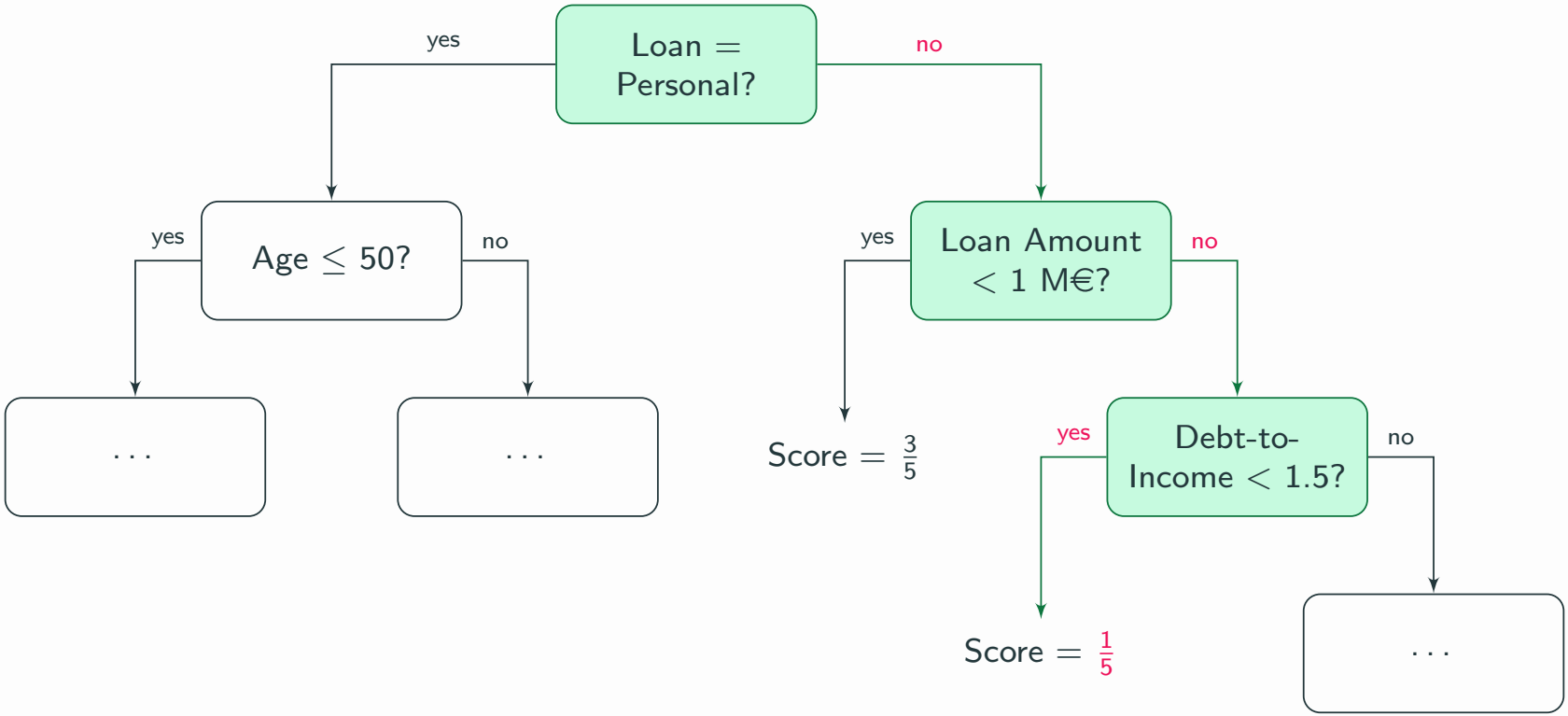
Regression Models: Illustration



Loan Eligibility

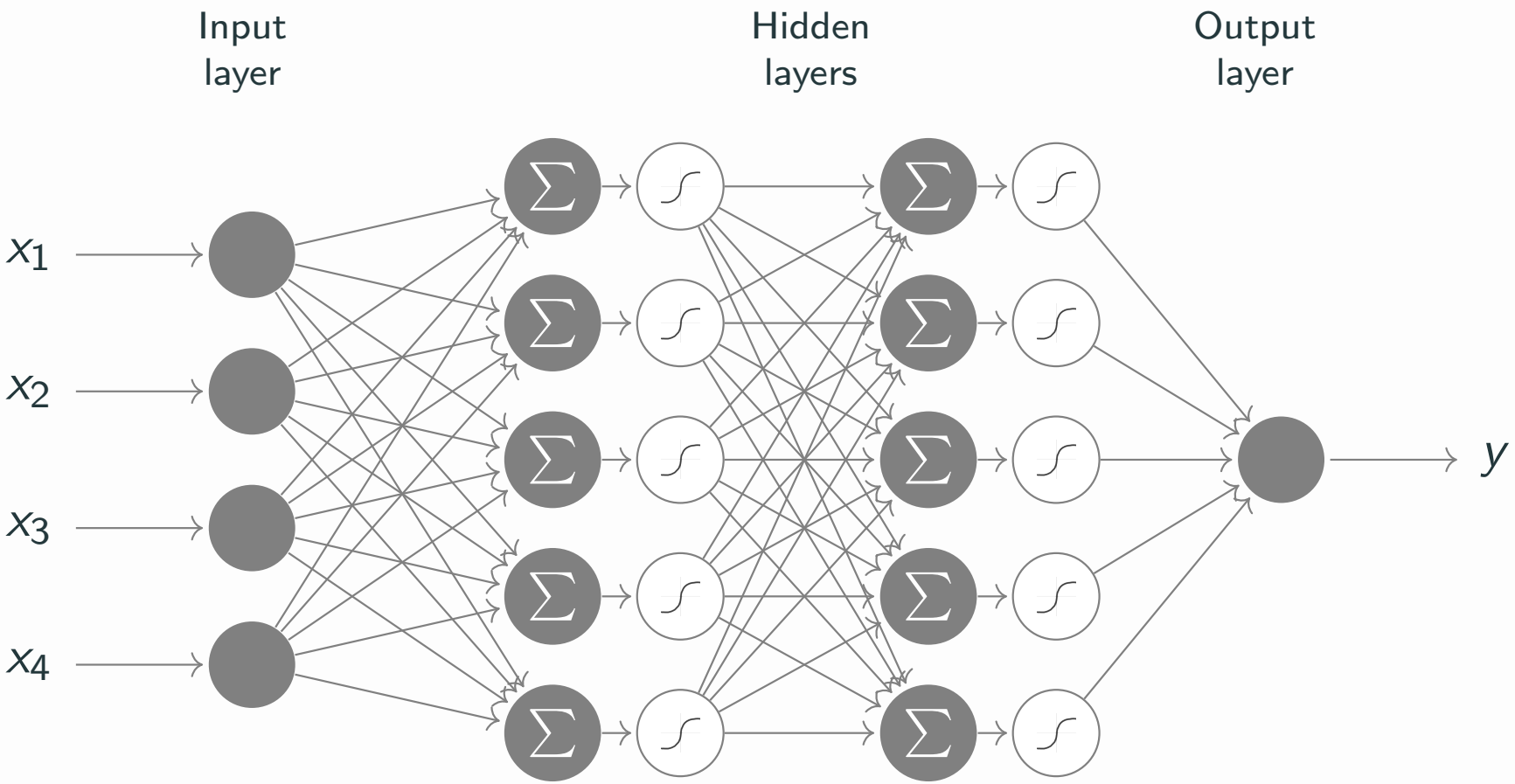
Each instance is a list of attributes about an applicant. The objective is to predict the applicant's likelihood of repaying the loan.

Regression Models: Interpretability



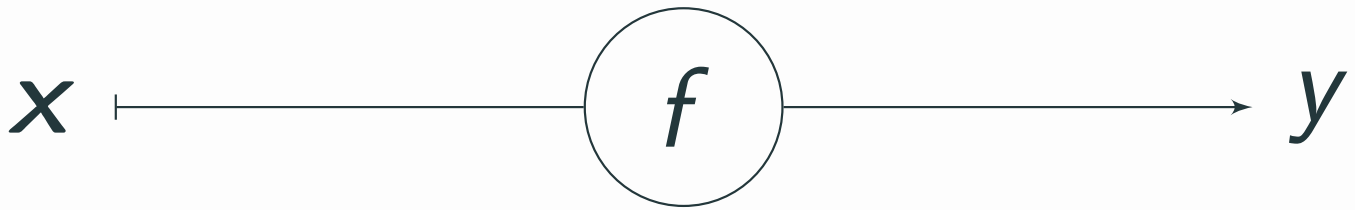
Although some regression models are interpretable . . .

Regression Models: ~~Interpretability~~



Most of them are not!

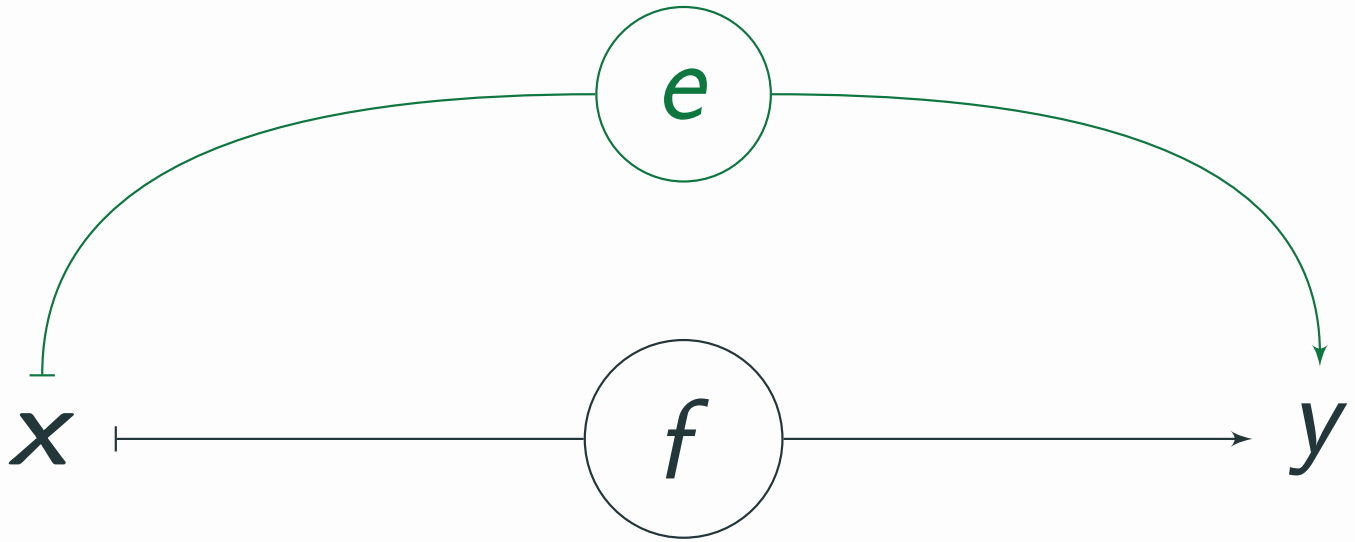
Explaining Regressors



Why y is the outcome of $\boldsymbol{x}$?

Thus, a key issue is to provide answers to why-questions.

Explaining Regressors



An explanation for an instance $\boldsymbol{x}$ with respect to a prediction model f is an interpretable surrogate model e that is consistent with f at $\boldsymbol{x}$.

Explaining Regressors: Setting

Black-Box Model

Any regression model f is viewed as a black box, with only access to the outcome of any queried instance.

Explaining Regressors: Setting

Black-Box Model

Any regression model f is viewed as a black box, with only access to the outcome of any queried instance.

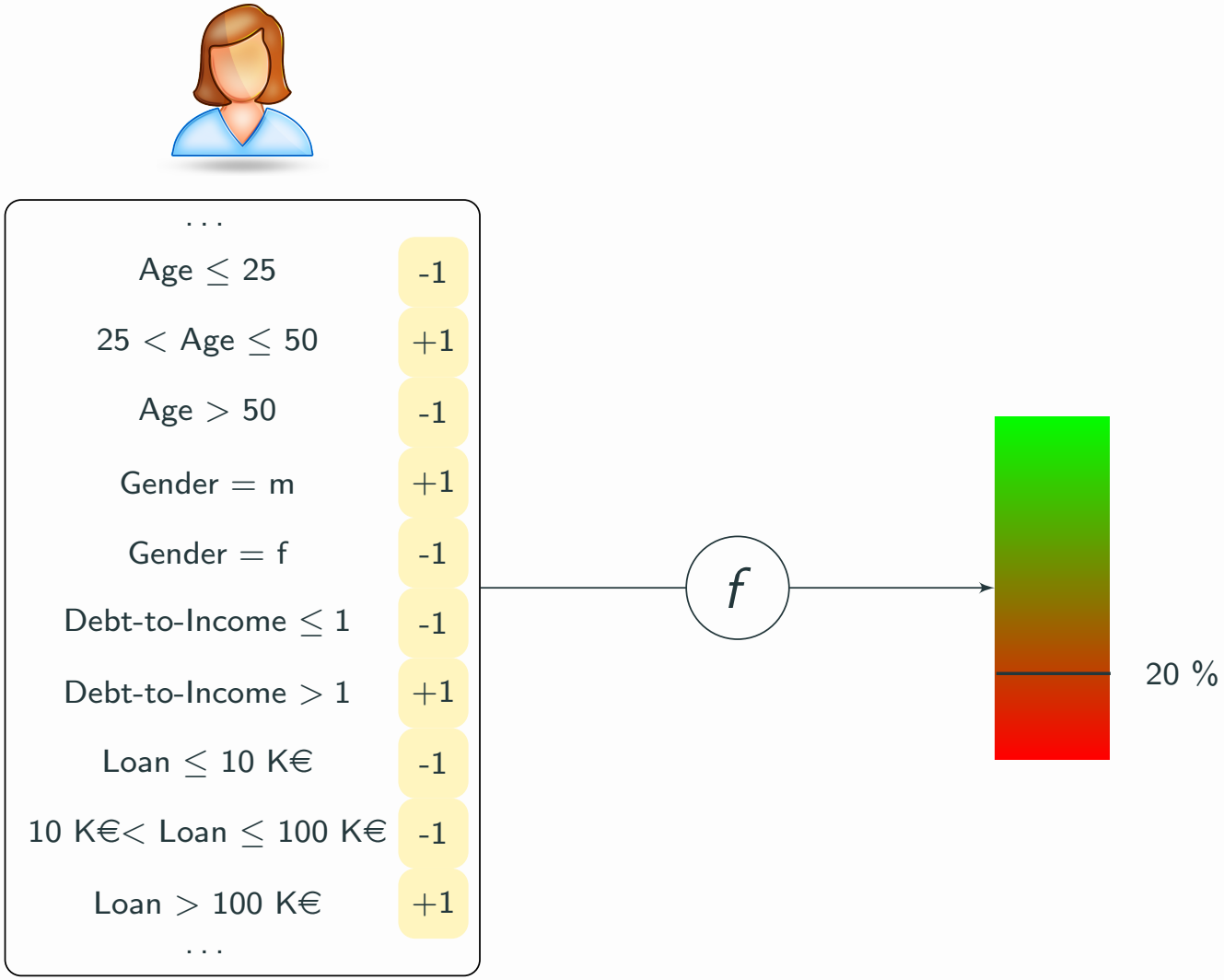
Main Assumption [Ribeiro et al., 2016]

The input space $\mathcal{X}$ is a Boolean hypercube, where each dimension is interpretable.

In other words, each instance is a vector $\boldsymbol{x} \in \{\pm 1\}^d$, where $[d] = \{1, \cdots, d\}$ is a set of interpretable features*.

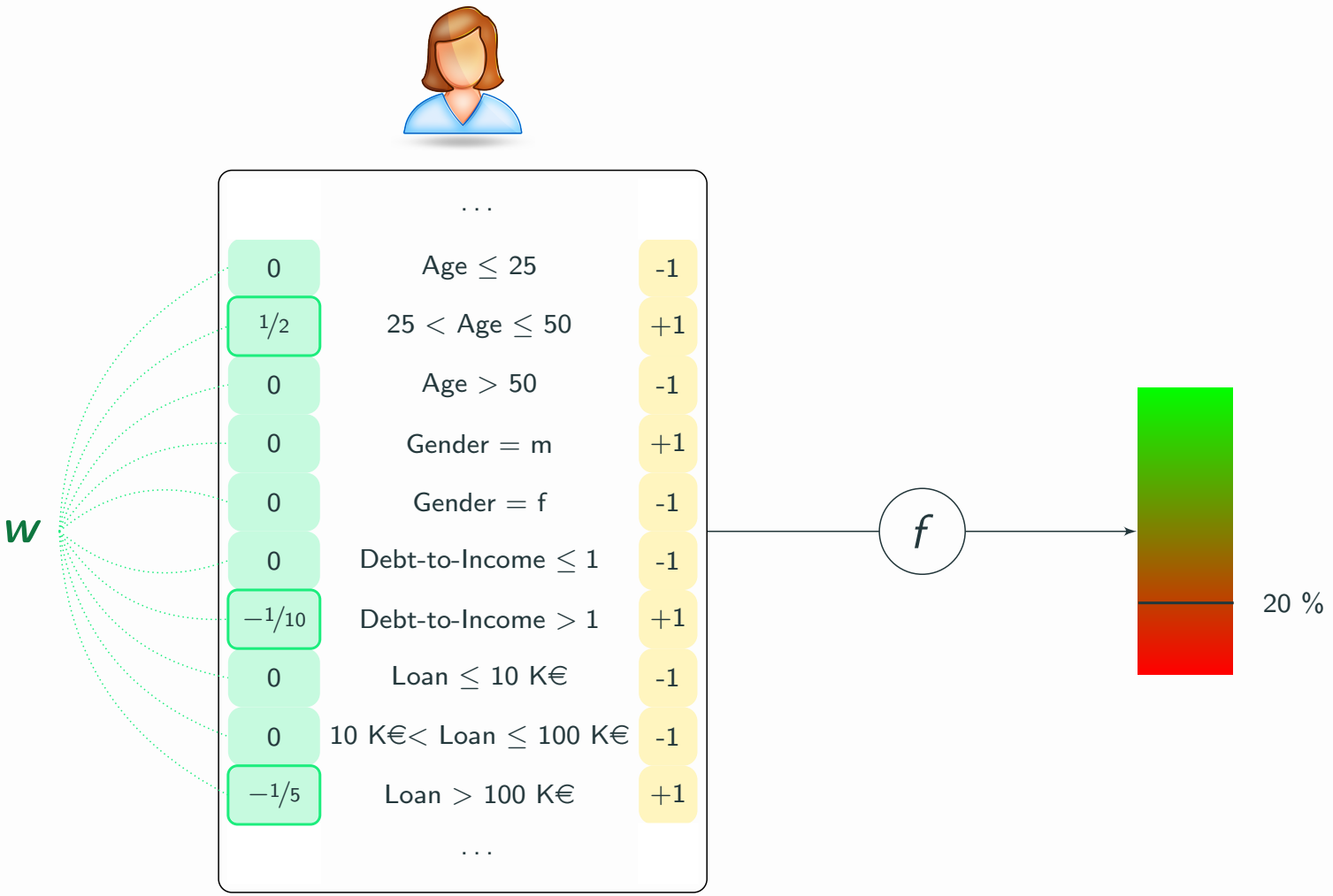
*Data instances over raw attributes are first transformed into Boolean vectors, using discretization techniques or interpretable latent spaces.

Explaining Regressors: Illustration



Using a set of interpretable features, let us explain why our applicant is not eligible for a loan ...

Explaining Regressors: Illustration



A clear way to explain the outcome of an instance is to identify a sparse vector of weights $\boldsymbol{w}$.

Explaining Regressors: Illustration

$$\tfrac{1}{2}\,[25 < \text{Age} \leq 50] - \tfrac{1}{10}\,[\text{DTI} > 1] - \tfrac{1}{5}\,[\text{Loan} > 100\ \text{K€}] \longrightarrow \text{Score} = \tfrac{1}{5}$$

This vector $\boldsymbol{w}$ can be seen as a weighted decision rule, where the head is determined by the sum of activated weights.

The explanation model e is therefore given by:

$$e(\boldsymbol{z}) = \boldsymbol{w} \cdot \boldsymbol{z} \quad \text{for all } \boldsymbol{z} \in \{\pm 1\}^d$$

Outline

What is a Good Explanation?

Two Main Criteria [Molnar, 2025]:

Conciseness

Is the explanation short enough to be understandable?

$\Longrightarrow$ Measured using $\|\boldsymbol{w}\|_0$, which is the number of nonzero weights of $\boldsymbol{w}$.

What is a Good Explanation?

Two Main Criteria [Molnar, 2025]:

Conciseness

Is the explanation short enough to be understandable?

$\Longrightarrow$ Measured using $\|\boldsymbol{w}\|_0$, which is the number of nonzero weights of $\boldsymbol{w}$.

Precision

Does the explanation predict the outcome as truthfully as possible?

$\Longrightarrow$ An explanation $\boldsymbol{w}$ is *sufficient* if $\boldsymbol{w} \cdot \boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x}$ implies $f(\boldsymbol{z}) = f(\boldsymbol{x})$ for all $\boldsymbol{z} \in \{\pm 1\}^d$.

What is a Good Explanation?

Two Main Criteria [Molnar, 2025]:

Conciseness

Is the explanation short enough to be understandable?

$\Longrightarrow$ Measured using $\|\boldsymbol{w}\|_0$, which is the number of nonzero weights of $\boldsymbol{w}$.

Precision

Does the explanation predict the outcome as truthfully as possible?

$\Longrightarrow$ An explanation $\boldsymbol{w}$ is *sufficient* if $\boldsymbol{w} \cdot \boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x}$ implies $f(\boldsymbol{z}) = f(\boldsymbol{x})$ for all $\boldsymbol{z} \in \{\pm 1\}^d$.

Unfortunately, both criteria are clashing: sufficient explanations may require too many weights to be understandable!

Probabilistic Explanations

Definition

Given a probability distribution $\mathcal{D}$ over $\{\pm 1\}^d$, the precision error of an explanation $\boldsymbol{w}$ for $\boldsymbol{x}$ is defined as

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}]$$

In other words, the error of $\boldsymbol{w}$ is the expected gap between $f(\boldsymbol{z})$ and $f(\boldsymbol{x})$, when the projections of $\boldsymbol{z}$ and $\boldsymbol{x}$ to $\boldsymbol{w}$ are the same.

Computing Probabilistic Explanations

Given

- a black-box model f,
- an instance $\boldsymbol{x}$ to explain,
- a probability distribution $\mathcal{D}$ over instances,
- a conciseness parameter k,

our problem is to

$$\begin{aligned} &\text{Minimize} && \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}] \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

Computing Probabilistic Explanations

Given

- a black-box model f,
- an instance $\boldsymbol{x}$ to explain,
- a probability distribution $\mathcal{D}$ over instances,
- a conciseness parameter k,

our problem is to

Objective Function

$$\begin{aligned} &\text{Minimize} && \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}] \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

Computing Probabilistic Explanations

Given

- a black-box model f,
- an instance $\boldsymbol{x}$ to explain,
- a probability distribution $\mathcal{D}$ over instances,
- a conciseness parameter k,

our problem is to

Objective Function

Minimize $\mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w} \cdot \boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x}]$

Subject to $\boldsymbol{w} \cdot \boldsymbol{x} = f(\boldsymbol{x})$ — Consistency Constraint

$\|\boldsymbol{w}\|_0 \leq k$ — Conciseness Constraint

Computing Probabilistic Explanations

Computational Complexity

The problem of finding an explanation $\boldsymbol{w}$ of size at most k and precision error at most ε is $\mathrm{NP}^{\mathrm{PP}}$-hard.

Thus, the problem involves *two independent* sources of complexity:

- Evaluating the precision error of a given vector $\boldsymbol{w}$ is PP-hard,
- Finding a vector $\boldsymbol{w}$ of size at most k that achieves minimal precision error is NP-hard.

Outline

Optimizing Precision

Consider again our optimization problem:

$$\begin{aligned} &\text{Minimize} && \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}] \quad \cdots\cdots\ P(\boldsymbol{w}) \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

The objective function $P(\boldsymbol{w})$ involves a conditional expectation, which is very difficult to evaluate.

Optimizing Fidelity

Now, consider the following variant:

$$\begin{aligned} &\text{Minimize} && \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[(\boldsymbol{w}\cdot\boldsymbol{z} - f(\boldsymbol{z}))^2] \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

$F(\boldsymbol{w})$

By substituting the precision error $P(\boldsymbol{w})$ with fidelity error $F(\boldsymbol{w})$ [Ribeiro et al., 2016], the objective function is an unconditional expectation that can be approximated through sampling.

Optimizing Empirical Fidelity

Finally, given a sample set $\{(\boldsymbol{z}_1, f(\boldsymbol{z}_1)), \cdots, (\boldsymbol{z}_m, f(\boldsymbol{z}_m))\}$ drawn from the distribution $\mathcal{D}$ and labeled by the predictor f, consider the problem:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \quad \cdots\cdots \hat{F}(\boldsymbol{w}) \\ \text{Subject to} \quad & \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

By approximating the fidelity $F(\boldsymbol{w})$ through empirical fidelity $\hat{F}(\boldsymbol{w})$, the objective function is now easy to evaluate.

Optimizing Empirical Fidelity

Approximation Guarantees

Let $\boldsymbol{w}^*$ be an optimal explanation for the precision. Then, using a number of samples m that is logarithmic in d and quadratic in k, any explanation $\boldsymbol{w}$ that is optimal for the empirical fidelity satisfies with high probability:

$$P(\boldsymbol{w}) \leq \sqrt{\hat{F}(\boldsymbol{w}^*)} + o(1)$$

Optimizing Empirical Fidelity

Approximation Guarantees

Let $\boldsymbol{w}^*$ be an optimal explanation for the precision. Then, using a number of samples m that is logarithmic in d and quadratic in k, any explanation $\boldsymbol{w}$ that is optimal for the empirical fidelity satisfies with high probability:

$$P(\boldsymbol{w}) \leq \sqrt{\hat{F}(\boldsymbol{w}^*)} + o(1)$$

Computing probabilistic explanations of optimal empirical fidelity can be solved via Mixed Integer Programming.

Outline

Optimizing Empirical Fidelity Remains Hard

Even with approximation guarantees, the final problem:

$$\begin{aligned} &\text{Minimize} && \tfrac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

Conciseness Constraint

is still NP-hard because the conciseness constraint is not convex!

Optimizing Empirical Fidelity Remains Hard

Even with approximation guarantees, the final problem:

$$\begin{aligned} &\text{Minimize} && \tfrac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \\ &\text{Subject to} && \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & && \|\boldsymbol{w}\|_0 \leq k \end{aligned}$$

Conciseness Constraint

is still NP-hard because the conciseness constraint is not convex!

$\Longrightarrow$ We thus need additional assumptions to achieve polynomial-time efficiency.

Uniform Distribution

Concentration Inequality [Achlioptas, 2001]

Let $\mathcal{D}$ be the uniform distribution over $\{\pm 1\}^d$. Then, for any $\boldsymbol{w} \in \mathbb{R}^d$ and any $\varepsilon \in (0, 1)$, we have

$$\mathbb{P}_{\boldsymbol{Z} \sim \mathcal{D}^m} \left[\left| \frac{1}{m} \|\boldsymbol{Z}\boldsymbol{w}\|_2^2 - \|\boldsymbol{w}\|_2^2 \right| > \varepsilon \right] \leq 2e^{-\Omega(m)}$$

Uniform Distribution

Concentration Inequality [Achlioptas, 2001]

Let $\mathcal{D}$ be the uniform distribution over $\{\pm 1\}^d$. Then, for any $\boldsymbol{w} \in \mathbb{R}^d$ and any $\varepsilon \in (0, 1)$, we have

$$\mathbb{P}_{\boldsymbol{Z} \sim \mathcal{D}^m} \left[\left| \frac{1}{m} \|\boldsymbol{Z}\boldsymbol{w}\|_2^2 - \|\boldsymbol{w}\|_2^2 \right| > \varepsilon \right] \leq 2e^{-\Omega(m)}$$

Thus, if the number of samples m is sufficiently large, any matrix $\boldsymbol{Z} \sim \mathcal{D}^m$ satisfies, with high probability, the Restricted Isometry Property for all k-sparse vectors $\boldsymbol{w}$ [Baraniuk et al., 2008]:

$$(1 - \beta_k)\|\boldsymbol{w}\|_2^2 \leq \frac{1}{m}\|\boldsymbol{Z}\boldsymbol{w}\|_2^2 \leq (1 + \beta_k)\|\boldsymbol{w}\|_2^2$$

Iterative Hard Thresholding

Input Query $(\boldsymbol{x}, f(\boldsymbol{x}))$, sparsity level k, samples $(\boldsymbol{Z}, \boldsymbol{y})$

Initialize $\boldsymbol{w}_0 = 0$

For each $t = 1, 2, \ldots$ do

$$\boldsymbol{v}_t = \boldsymbol{w}_{t-1} - \tfrac{1}{m}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{w}_{t-1} - \boldsymbol{y})$$

Gradient Descent

$$\boldsymbol{w}_t = \operatorname{argmin}_{\boldsymbol{w} \in \mathcal{W}} \|\boldsymbol{v}_t - \boldsymbol{w}\|_2$$

Projection onto feasible explanations

$$\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} \cdot \boldsymbol{x} = f(\boldsymbol{x}) \text{ and } \|\boldsymbol{w}\|_0 \leq k\}$$

Optimizing Empirical Fidelity under Uniform Distribution

Efficiency Guarantees

Let $\boldsymbol{w}^*$ be an optimal explanation for the precision. Then, using a polynomial number of samples drawn uniformly at random, the IHT algorithm is guaranteed to find, with high probability, a k-sparse explanation $\boldsymbol{w}_t$ that achieves

$$P(\boldsymbol{w}_t) \leq 7\sqrt{\hat{F}(\boldsymbol{w}^*)} + o(1)$$

Furthermore, $\boldsymbol{w}_t$ can be computed in polynomial time with respect to d, k, and $\log_2\lceil 1/\hat{F}(\boldsymbol{w}^*)\rceil$.

Outline

Setup

Benchmarks

- Prediction Tasks: 20 regression datasets from `OpenML`
- Black-box f: neural networks (MLP) learned from train set
- Instance $\boldsymbol{x}$: selected uniformly at random from the test set
- Distributions: $\mathcal{D}$ parameterized by spread $\sigma \in [0, 1]$
- Explanation Sizes: $k \in \{1, \cdots, 10\}$
- Number of samples: $m = 1000$
- Timeout: 60 seconds

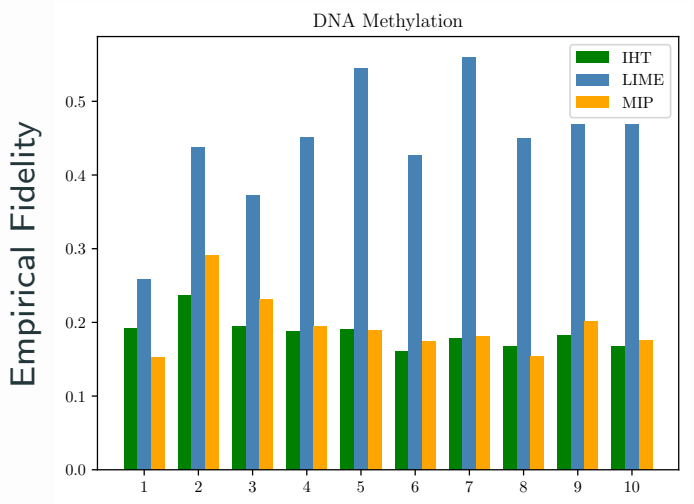
Competitors

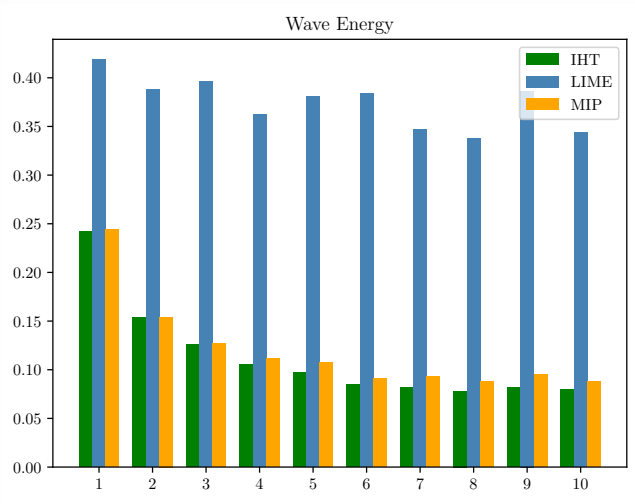
- IHT, MIP (Gurobi solver) versus CVX (convex relaxation), LIME, and MAPLE.

Results

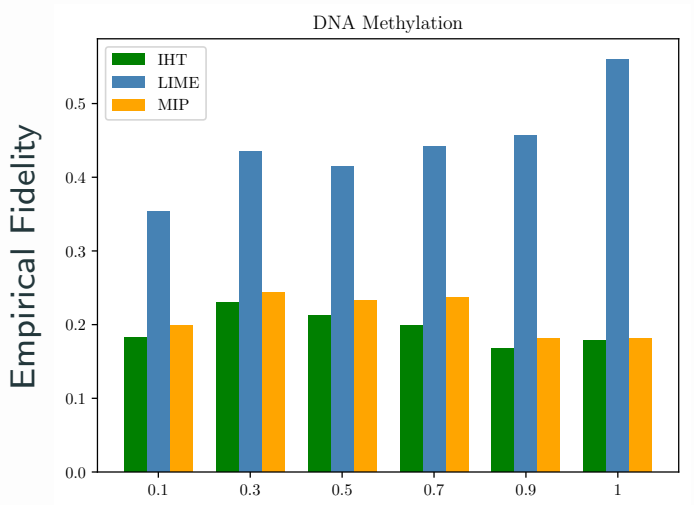
Benchmark	Empirical Fidelity				
Name	CVX	IHT	LIME	MAPLE	MIP
Airfoil Self Noise	0.040	0.055	0.321	0.218	**0.049**
Auto MPG	0.031	0.069	0.338	0.122	**0.039**
Liver Disorders	0.059	0.091	0.209	0.147	**0.068**
Medical Charges	0.040	**0.049**	0.408	0.204	**0.049**
Ailerons	0.050	0.201	0.647	0.113	**0.085**
Auto Imports	0.067	0.232	0.528	0.148	**0.107**
DNA Methylation	0.121	**0.192**	0.582	0.168	**0.191**
NCI 60 Thioguanine	0.062	0.235	0.534	0.108	**0.132**
Student Performance	0.074	0.143	0.454	0.169	**0.105**
Wave Energy	0.017	**0.080**	0.301	0.128	0.091

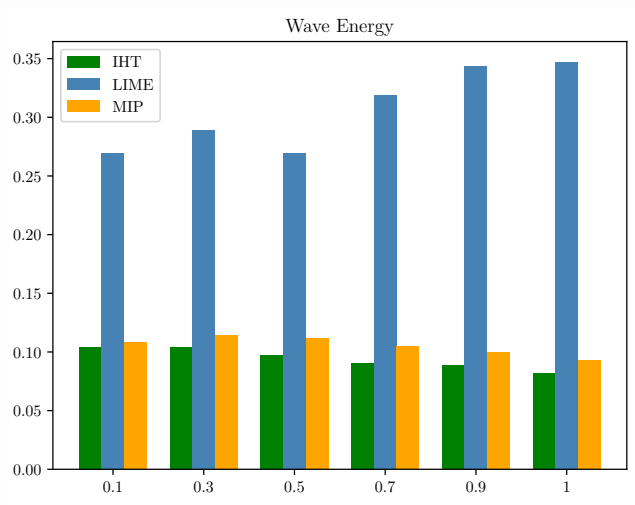
Results on 4 low-dimensional benchmarks and 6 medium-dimensional benchmarks, using $k = 7$, $m = 1000$, and $\sigma = 1$. Entries highlighted in green indicate that all generated explanations were k-sparse.





Increasing Explanation Sizes k





Increasing Spread σ

Comparisons among IHT, Lime and MIP for several parameters.

Conclusion

- Probabilistic explanations achieve a balance between *conciseness* and *precision*.
- However, computing these explanations is very challenging (NP^{PP}-hard).
- By replacing precision with fidelity, they can be approached with Mixed Integer Programming, while offering approximation guarantees.
- Under the uniform distribution, these explanations can be efficiently approached using Iterative Hard Thresholding.
- Empirical results on real-world benchmarks support our theoretical findings.

Conclusion

- Probabilistic explanations achieve a balance between *conciseness* and *precision*.
- However, computing these explanations is very challenging (NP^{PP}-hard).
- By replacing precision with fidelity, they can be approached with Mixed Integer Programming, while offering approximation guarantees.
- Under the uniform distribution, these explanations can be efficiently approached using Iterative Hard Thresholding.
- Empirical results on real-world benchmarks support our theoretical findings.

Thank you for your attention!

Bibliography

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the 20th ACM Symposium on Principles of Database Systems (PODS)*, page 274–281, 2001.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28:253–263, 2008.

Christoph Molnar. *Interpretable Machine Learning*. leanpub.com, 3rd edition, 2025. URL `https://christophm.github.io/interpretable-ml-book`.

Marco Túlio Ribeiro, Sameer Singh, and Carlos 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.