Probabilistic Explanations for Regression Models

Frederic Koriche[1] **Jean-Marie Lagniez**[1] **Chi Tran**[1]

[1]Univ. Artois, CNRS, Centre de Recherche en Informatique de Lens (CRIL), France

Abstract

Formal explainability is an emerging field that aims to provide mathematically guaranteed explanations for the predictions made by machine learning models. Recent work in this area focuses on computing "probabilistic explanations" for the predictions made by classifiers based on specific data instances. The goal of this paper is to extend the concept of probabilistic explanations to the regression setting, treating the target regressor as a black box function. The class of probabilistic explanations consists of linear functions that meet a sparsity constraint, alongside a hyperplane constraint defined for the data instance being explained. While minimizing the precision error of such explanations is generally $\mathsf{NP}^{\mathsf{PP}}$-hard, we demonstrate that it can be approximated by substituting the precision measure with a fidelity measure. Optimal explanations based on this fidelity objective can be effectively approached using Mixed Integer Programming (MIP). Moreover, we show that for certain distributions used to define the precision measure, explanations with approximation guarantees can be computed in polynomial time using a variant of Iterative Hard Thresholding (IHT). Experiments conducted on various datasets indicate that both the MIP and IHT approaches outperform the state-of-the-art LIME and MAPLE explainers.

1 INTRODUCTION

As machine learning models increasingly impact critical decisions in areas such as criminal justice, medical diagnosis, and social scoring, the significance of ethics, fairness, and safety in these models has become more apparent than ever. In response to this need, Explainable Artificial Intelligence (XAI) has developed a range of explanation techniques that help users understand these models without requiring in-depth knowledge of their inner workings [Miller et al., 2022, Molnar, 2022]. Recently, the field of *formal explainability* has emerged as a promising subdiscipline, concentrating on providing explanations with mathematical guarantees concerning quality, size, and semantics [Ignatiev, 2020, Marques-Silva and Ignatiev, 2022]. The aim of formal explainability is to establish theoretical foundations for explaining predictions made by machine learning models, so as to calibrate trust and confidence in their capabilities.

A well-studied problem in formal explainability is to identify a rule that explains *why* a given data instance $\boldsymbol{x}$ is classified as $f(\boldsymbol{x})$ by a classifier f. This rule can be described as a subset S of features, such that any change in the values of features outside S does not affect the outcome $f(\boldsymbol{x})$. Since the restriction of $\boldsymbol{x}$ to S, denoted by $\boldsymbol{x}_S$, contains enough information to determine $f(\boldsymbol{x})$, the feature subset S is often referred to as a (weak) *abductive explanation* [Cooper and Marques-Silva, 2023], also called *sufficient reason* [Darwiche and Hirth, 2020]. However, despite the appealing soundness of abductive explanations, their size often exceeds the cognitive limits of human users. As suggested by Miller [1956], our ability to reason about multiple features is typically limited to seven, plus or minus two elements. This limitation has been reinforced by numerous cognitive science experiments (see e.g. [Saaty and Özdemir, 2003]), and empirical research in XAI indicates that explanations should be concise [Lage et al., 2019].

Therefore, achieving a balance between precision and conciseness is crucial when generating explanations for predictive models. The concept of *probabilistic explanations* [Wäldchen et al., 2021, Izza et al., 2023] embodies this balance. In this context, the *precision error* of a feature set S is the probability that f separates a random instance $\boldsymbol{z}$ from $\boldsymbol{x}$, when the restrictions of $\boldsymbol{z}$ and $\boldsymbol{x}$ to S are indistinguishable. The precision error is evaluated according to a predefined distribution, such as the uniform distribution over all data instances, or some neighborhood distribution centered at $\boldsymbol{x}$. Based on this measure, the computation of

probabilistic explanations can be framed as a constrained stochastic optimization problem. For example, if we aim to find an explanation with the lowest precision error under a user-supplied size limit k, the task is to

$$\begin{array}{ll} \text{minimize} & \mathbb{P}_{\boldsymbol{z}}[f(\boldsymbol{z}) \neq f(\boldsymbol{x}) \mid \boldsymbol{z}_S = \boldsymbol{x}_S] \\ \text{subject to} & |S| \leq k \end{array} \tag{P1}$$

To the best of our knowledge, probabilistic explanations have mostly been studied within the context of classification. However, considering the variety of available regression models, a logical question arises: *how can probabilistic explanations be extended to the regression setting?*

This paper addresses the above question without making assumptions about the structure of the regression model f. For instance, f could be represented by a tree ensemble, a support vector machine, or a deep neural network. In our algorithms, f is treated as a black-box function.

When explaining the prediction $f(\boldsymbol{x})$ made by f for a data instance $\boldsymbol{x}$, feature subsets alone are often insufficient to describe the relationship between input features and continuous output values. Therefore, this study considers explanations in the form of *linear models* $\boldsymbol{w}$ that satisfy the hyperplane condition $\boldsymbol{w} \cdot \boldsymbol{x} = f(\boldsymbol{x})$. Such a constraint ensures that any explanation $\boldsymbol{w}$ is consistent with f at $\boldsymbol{x}$. The conciseness or *sparsity* of $\boldsymbol{w}$ is measured by the number of its nonzero coefficients, denoted as $\|\boldsymbol{w}\|_0$.

To quantify how "sufficient" a sparse linear model is in determining a regression model with adequate precision, we replace the conditional zero-one loss function in (P1) with a conditional absolute loss function. Thus, the precision error of an explanation $\boldsymbol{w}$ for the value $f(\boldsymbol{x})$ of some data instance $\boldsymbol{x}$ is defined by the conditional expected loss of $|f(\boldsymbol{z}) - f(\boldsymbol{x})|$ given that $\boldsymbol{w} \cdot \boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x}$. Again, the precision is evaluated according to some predefined distribution over the instance space. With these notions in hand, the problem examined in this paper is to

$$\begin{array}{ll} \text{minimize} & \mathbb{E}_{\boldsymbol{z}}[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w} \cdot \boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x}] \\ \text{subject to} & \boldsymbol{w} \cdot \boldsymbol{x} = f(\boldsymbol{x}) \text{ and } \|\boldsymbol{w}\|_0 \leq k \end{array} \tag{P2}$$

In Section 4, we show that when f is represented by a neural network, (P2) is hard for $\mathsf{NP}^{\mathsf{PP}}$, a complexity class that is beyond the capabilities of modern solvers. However, this hardness result does not preclude the existence of algorithms that offer *additive* approximation guarantees on the conciseness and the precision of optimal explanations.

In Section 5, we show that the precision error of feasible solutions in (P2) is upper-bounded by their *fidelity error*, a measure often used in model-agnostic explainability [Li et al., 2021]. By replacing the objective in (P2) with the empirical fidelity error, the corresponding problem becomes:

$$\begin{array}{ll} \text{minimize} & \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \\ \text{subject to} & \boldsymbol{w} \cdot \boldsymbol{x} = f(\boldsymbol{x}) \text{ and } \|\boldsymbol{w}\|_0 \leq k \end{array} \tag{P3}$$

This formulation, which involves a non-conditional expected loss function as the objective, is a variant of the well-studied *sparse regression* problem [Natarajan, 1995]. While this problem remains NP-hard, it can be approached using *Mixed Integer Programming* (MIP) with a polynomial number of queries to f. The corresponding explanations are k-sparse, and with high probability, their precision error is at most $\sqrt{\gamma^*} + o(1)$, where γ^* is the optimal value of (P3).

In Section 6, we present a variant of the *Iterative Hard Thresholding* (IHT) algorithm [Blumensath and Davies, 2009, Garg and Khandekar, 2009] that computes approximate solutions to (P3) in polynomial time. For the uniform distribution, these explanations are k-sparse, and with high probability, their precision error is at most $7\sqrt{\gamma^*} + o(1)$.

From an empirical standpoint, we compare in Section 7 the MIP and IHT approaches with the popular LIME [Ribeiro et al., 2016] and MAPLE [Plumb et al., 2018] explainers. Through experiments on various datasets, we demonstrate that both MIP and IHT approaches outperform these state-of-the-art explainers in terms of fidelity while using a reasonable amount of time for the MIP solver.

2 RELATED WORK

Probabilistic explanations have gained increasing attention in the field of formal explainability due to their flexibility. As outlined in (P1), we can set a sparsity level k and request a feature subset S of size at most k that minimizes precision error [Koriche et al., 2024]. Alternatively, we can fix a precision level ϵ and ask for a smallest feature subset S with an error of at most $1 - \epsilon$ [Izza et al., 2023]. However, this flexibility comes at a cost: Wäldchen et al. [2021] demonstrated that deciding whether there exists a k-sparse ϵ-precise explanation S for the prediction $f(\boldsymbol{x})$ made by a neural network f on a data instance $\boldsymbol{x}$ is a $\mathsf{NP}^{\mathsf{PP}}$-hard problem. Additionally, they showed that minimizing the size of an ϵ-precise explanation is NP-hard to approximate within a factor of $d^{1-\delta}$ for any $\delta > 0$, where d is the dimension of x.

For these reasons, the tractability and approximability of probabilistic explanations have been explored for simpler classifiers, such as decision trees [Arenas et al., 2022, Bounia and Koriche, 2023] and linear threshold functions [Subercaseaux et al., 2025]. In cases where f is a black-box classifier, Blanc et al. [2021] demonstrated that if the instance $\boldsymbol{x}$ being explained is drawn from a uniform distribution, then with high probability, an ϵ-precise explanation S of size k' can be derived from the path $T(\boldsymbol{x})$ of a depth-k' decision

tree T with fidelity error $\mathbb{P}_{\boldsymbol{z}}[T(\boldsymbol{z}) \neq f(\boldsymbol{z})] \leq \epsilon$. When k' is polynomial in the "average certificate complexity" of f, the decision tree T can be implicitly learned in polynomial time. While our approach for the regression setting shares some similarities with their findings, we do not assume that $\boldsymbol{x}$ is selected uniformly at random.

In a broader context, various model-agnostic methods have been proposed to extrapolate a linear explanations from the neighborhood of data instances [Ribeiro et al., 2016, Plumb et al., 2018, Agarwal et al., 2021, Zhao et al., 2021]. A common goal is to minimize the unconstrained objective

$$\frac{1}{m}\sum_{i=1}^{m} \phi_{\boldsymbol{x}}(\boldsymbol{z}_i)(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 + \psi(\boldsymbol{w})$$

Here, $\{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$ is a set of labeled samples generated from some neighborhood distribution around $\boldsymbol{x}$, $\phi_{\boldsymbol{x}}(\boldsymbol{z}_i)$ assesses the importance of $\boldsymbol{z}_i$, and $\psi(\boldsymbol{w})$ penalizes the complexity of $\boldsymbol{w}$. For example, in the LIME method [Ribeiro et al., 2016], $\phi_{\boldsymbol{x}}(\boldsymbol{z}_i)$ is a normalized distance between $\boldsymbol{z}_i$ and $\boldsymbol{x}$, while in the MAPLE method [Plumb et al., 2018], $\phi_{\boldsymbol{x}}(\boldsymbol{z}_i)$ measures the average number of times $\boldsymbol{z}_i$ ends up in the same leaf as $\boldsymbol{x}$ in a random forest trained from f. Despite their popularity, these heuristic methods do not always provide theoretical guarantees regarding the consistency, fidelity, or sparsity of extrapolated explanations. This contrasts with our MIP and IHT approaches, which aim to solve (P3), incorporating consistency and sparsity as constraints, and defining fidelity as the objective.

Notation. Plain letters represent functions and scalars, while boldface letters represent vectors and matrices. The all-ones vector is denoted as $\mathbf{1}$ and the all-zeros vector as $\mathbf{0}$. For a positive integer d, we use $[d]$ to denote the set $\{1, \ldots, d\}$. Additionally, we use $\mathbf{1}_S$ to denote the indicator vector in $\{0,1\}^d$ of a subset $S \subseteq [d]$, and we use $\mathbb{1}[E]$ to denote the indicator function in $\{0,1\}$ of an event $E \subseteq \{0,1\}^d$. The support set of a vector $\boldsymbol{w} \in \mathbb{R}^d$, denoted as $\text{support}(\boldsymbol{w})$, is the set of coordinates $j \in [d]$ for which $w_j \neq 0$. The scalar product of two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ is denoted as $\boldsymbol{v}\cdot\boldsymbol{w}$, and the coordinate-wise (or Hadamard) product is denoted as $\boldsymbol{v}\odot\boldsymbol{w}$. For a scalar $p \in [0,\infty]$, the L_p norm of $\boldsymbol{w}$ is denoted as $\|\boldsymbol{w}\|_p$. The limit cases are $\|\boldsymbol{w}\|_0 = |(\text{support}(\boldsymbol{w}))|$ and $\|\boldsymbol{w}\|_\infty = \max_{j=1}^d |w_j|$. For a scalar $r \geq 0$, the L_p ball of radius r is defined as

$$\mathcal{B}_p(r) = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_p \leq r\}$$

For a vector $\boldsymbol{u} \in \mathbb{R}^d$ and a scalar $r \in \mathbb{R}$, the hyperplane at $(\boldsymbol{u}, r)$ is defined as

$$\mathcal{H}(\boldsymbol{u}, r) = \{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{u}\cdot\boldsymbol{w} = r\}$$

Finally, the Euclidean projection of a vector $\boldsymbol{w} \in \mathbb{R}^d$ onto a set $\mathcal{U} \subseteq \mathbb{R}^d$ is given by

$$\Pi_{\mathcal{U}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{w} - \boldsymbol{u}\|_2.$$

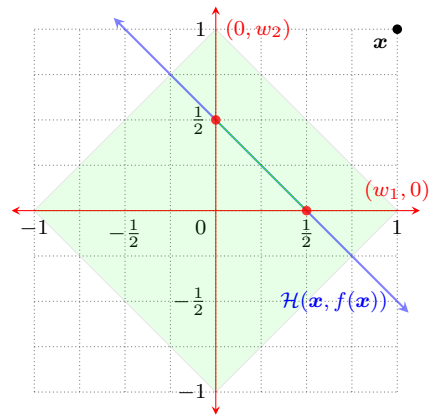

Figure 1: A geometric illustration of 1-sparse linear explanations $\boldsymbol{w}$, where $\boldsymbol{x} = (1,1)$ and $f(\boldsymbol{x}) = \frac{1}{2}$. The hyperplane $\mathcal{H}(\boldsymbol{x}, f(\boldsymbol{x}))$ is shown in blue, while the L_0 ball $\mathcal{B}_0(1)$ is depicted in red. The intersection of these two elements is represented by two red points. The convex hull of this intersection forms the green segment, with the lozenge highlighting the L_1 ball $\mathcal{B}_1(1)$.

3 PROBLEM FORMULATION

In this study, we consider explanation tasks where data instances are defined over a set of *interpretable literals*. For instance, consider a bank customer wanting to understand why her loan application received a score of $-\frac{1}{4}$, which is below the acceptance threshold. Interpretable literals such as [Income $\geq$ 70K\$], [Debt-To-Income (DTI) ratio $\leq$ 35%], and [Proof of Address = Yes] could be utilized. A clear and concise explanation could be provided using an if-then rule with weighted features, such as

$$\tfrac{1}{2}[\text{Income} \geq 70\text{K\$}] - \tfrac{3}{4}[\text{DTI ratio} > 35\%] \rightarrow \text{Score} = -\tfrac{1}{4}$$

More formally, let $[d]$ denote the set of interpretable literals. By treating these literals as binary features, the regression models explored in this study are pseudo-Boolean functions of the form $f : \{\pm 1\}^d \rightarrow [-1, +1]$.[1] Here, any input to f is a data instance $\boldsymbol{x} \in \{\pm 1\}^d$, where x_j indicates whether the jth literal occurs positively or negatively in $\boldsymbol{x}$.

A *linear explanation* for $f(\boldsymbol{x})$ is a vector $\boldsymbol{w} \in \mathbb{R}^d$ that satisfies the equation $\boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x})$. As illustrated in the previous example, such an explanation can be interpreted as an if-then rule over weighted literals: the head corresponds to $f(\boldsymbol{x})$, and the body consists of pairs (j, w_j) for which $x_j w_j \neq 0$. An explanation $\boldsymbol{w}$ is *k-sparse* if $\|\boldsymbol{w}\|_0 \leq k$. As illustrated in Figure 1, the set of k-sparse explanations for $f(\boldsymbol{x})$ is formed by the intersection of two objects: the hyperplane $\mathcal{H}(\boldsymbol{x}, f(\boldsymbol{x}))$ and the L_0 ball $\mathcal{B}_0(k)$. While the former is convex, the latter is not.

[1] Our theoretical results can easily be extended to co-domains $[-c, +c]$, provided that c is constant.

The quality of probabilistic explanations is assessed in relation to a probability distribution $\mathcal{D}$ over $\{\pm 1\}^d$. For instance, $\mathcal{D}$ could represent the uniform distribution $\mathcal{U}$ across $\{\pm 1\}^d$ or, more restrictively, a neighborhood distribution surrounding the instance that is being explained. The *precision error* of a vector $\boldsymbol{w} \in \mathbb{R}^d$ with respect to a model f, a data instance $\boldsymbol{x}$, and a distribution $\mathcal{D}$ is defined as follows:

$$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[|f(\boldsymbol{z}) - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}\right] \quad (1)$$

In other words, the precision of $\boldsymbol{w}$ measures the discrepancy between $f(\boldsymbol{z})$ and $f(\boldsymbol{x})$ for random instances $\boldsymbol{z}$ that are aligned in the same direction as $\boldsymbol{x}$ in relation to $\boldsymbol{w}$.

With these concepts in mind, the decision version of (P2) is referred to as the SPARSE LINEAR EXPLANATION (SLE) problem, and formulated as follows:

Instance: A regression model $f : \{\pm 1\}^d \to [-1, +1]$, a data instance $\boldsymbol{x} \in \{\pm 1\}^d$, a probability distribution $\mathcal{D}$ over $\{\pm 1\}^d$, a sparsity level $k \geq 1$, and a precision parameter $\epsilon > 0$.

Question: Does there exist a linear explanation $\boldsymbol{w} \in \mathbb{R}^d$ for $f(\boldsymbol{x})$ such that $\|\boldsymbol{w}\|_0 \leq k$ and $\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq \epsilon$?

4 PROBLEM COMPLEXITY

To establish the computational hardness of our problem, we need a representation of f that allows us to evaluate its description length. To this end, we consider the class $\mathcal{N}$ of feedforward neural networks, which have weights and biases in $[-1, +1]$ and activation functions in $\{\sigma_{\text{LIN}}, \sigma_{\text{ReLU}}\}$, where $\sigma_{\text{LIN}}(u) = u$ and $\sigma_{\text{ReLU}}(u) = \max\{0, u\}$. The description length of f is determined by the number of gates in its representation. Additionally, we assume that the distribution $\mathcal{D}$ has a closed-form expression, allowing us to evaluate the probability $\mathcal{D}(\boldsymbol{z})$ of any $\boldsymbol{z}$ in polynomial time relative to the input dimension.

Theorem 1. For the representation class $\mathcal{N}$, the SPARSE LINEAR EXPLANATION problem is $\mathsf{NP}^{\mathsf{PP}}$-hard.

Proof. Consider the decision version of (P1), referred to as SPARSE SUBSET EXPLANATION (SSE). An instance of this problem is a tuple $I_{\text{SSE}} = (f, \boldsymbol{x}, k, \epsilon)$ such that $f : \{\pm 1\}^d \to \{\pm 1\}$ is a Boolean function represented by a Boolean circuit, $\boldsymbol{x} \in \{\pm 1\}^d$ is a data instance, and k and ϵ are parameters in $\mathbb{N}$ and $(0, 1]$, respectively. The goal is to decide whether there exists a subset $S \subseteq [d]$ of size at most k, such that $\mathsf{Q}_{f,\boldsymbol{x}}(S) \leq \epsilon$, where

$$\mathsf{Q}_{f,\boldsymbol{x}}(S) = \mathbb{P}_{\boldsymbol{z}\sim\mathcal{U}}\left[f(\boldsymbol{z}) \neq f(\boldsymbol{x}) \mid \boldsymbol{z}\odot\mathbf{1}_S = \boldsymbol{x}\odot\mathbf{1}_S\right]$$

Using $\delta = 1 - \epsilon$, any such subset S is called *δ-relevant subset* in [Wäldchen et al., 2021, Izza et al., 2023].

From an instance $I_{\text{SSE}} = (f, \boldsymbol{x}, k, \epsilon)$, we build an instance $I_{\text{SLE}} = (f', \boldsymbol{x}', \mathcal{D}, k', \epsilon')$ of our problem, defined as follows. Let $k' = k+1$, let $\epsilon' = \epsilon$ and let $\boldsymbol{x}' = (\boldsymbol{x}, 1)$. In addition, let $\mathcal{D}$ be the distribution over $\{\pm 1\}^{d+1}$ defined as $\mathcal{D}(\boldsymbol{z}, 1) = \mathcal{U}(\boldsymbol{z})$ and $\mathcal{D}(\boldsymbol{z}, -1) = 0$ for any $\boldsymbol{z} \in \{\pm 1\}^d$. Finally, let $f' : \{\pm 1\}^{d+1} \to [-1, +1]$ be the function:

$$f'(\boldsymbol{z}, -1) = f'(\boldsymbol{z}, 1) = \tfrac{1}{2}|f(\boldsymbol{x}) - f(\boldsymbol{z})|, \text{ for all } \boldsymbol{z} \in \{\pm 1\}^d$$

As shown in [Wäldchen et al., 2021], any Boolean circuit can be efficiently transformed into an equivalent neural network with integer weights and biases in $\{-1, 0, +1\}$, and activation functions in $\{\sigma_{\text{LIN}}, \sigma_{\text{ReLU}}\}$. Consequently, a representation in $\mathcal{N}$ for f' can be constructed in polynomial time from the neural representation of f, by simply adding the following units to its output:

$$\tfrac{1}{2}\sigma_{\text{LIN}}(\sigma_{\text{ReLU}}(f(\boldsymbol{x}) - f(\boldsymbol{z})), \sigma_{\text{ReLU}}(f(\boldsymbol{z}) - f(\boldsymbol{x})))$$

For a subset S, let $\boldsymbol{w} = (\mathbf{1}_S \odot \boldsymbol{x}, -|S|)$ denote the corresponding linear function. Since $(\mathbf{1}_S \odot \boldsymbol{x}) \cdot \boldsymbol{x} = |S|$, we know that $\boldsymbol{w}$ is a k-sparse explanation for $f'(\boldsymbol{x}')$. Furthermore, for any $\boldsymbol{z} \in \{\pm 1\}^d$, we have $\boldsymbol{w} \cdot (\boldsymbol{z}, 1) = \boldsymbol{w} \cdot (\boldsymbol{x}, 1)$ if and only if $(\mathbf{1}_S \odot \boldsymbol{x}) \cdot \boldsymbol{z} = |S|$, which is equivalent to $\mathbf{1}_S \odot \boldsymbol{x} = \mathbf{1}_S \odot \boldsymbol{z}$. This, together with the fact that $|f'(\boldsymbol{z}, 1) - f'(\boldsymbol{x}, 1)| = \frac{1}{2}|f(\boldsymbol{z}) - f(\boldsymbol{x})| = \mathbb{1}[f(\boldsymbol{z}) \neq f(\boldsymbol{x})]$ implies that $\mathsf{P}_{f',\boldsymbol{x}',\mathcal{D}}(\boldsymbol{w}) = \mathsf{Q}_{f,\boldsymbol{x}}(S)$. Therefore, S is a solution to I_{SSE} if and only if $\boldsymbol{w}$ is a solution to I_{SLE}. Since SSE is $\mathsf{NP}^{\mathsf{PP}}$-hard [Wäldchen et al., 2021, Theorem 2.4], it follows that SLE is $\mathsf{NP}^{\mathsf{PP}}$-hard. □

5 DEALING WITH PP-HARDNESS

Theorem 1 reveals that the problem of finding k-sparse linear explanations with a precision error of at most ϵ involves two independent sources of complexity. The first source, related to the NP-hardness of the problem, arises from the challenge of exploring all candidate support sets $S \subseteq [d]$ of size at most k and determining whether there exists an ϵ-precise linear explanation $\boldsymbol{w}$ with support S. The second source of complexity comes from the inherent difficulty in checking whether the precision error of $\boldsymbol{w}$ is indeed at most ϵ, which is itself a PP-hard problem.

In this section, we focus on the second source of complexity. The idea is to replace the precision error with the *fidelity error*, which serves as a surrogate function:

$$\mathsf{F}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[(\boldsymbol{w}\cdot\boldsymbol{z} - f(\boldsymbol{z}))^2\right] \quad (2)$$

Lemma 1. Let $f : \{\pm 1\}^d \to [-1, +1]$ be a regression model, let $\boldsymbol{x} \in \{\pm 1\}^d$ be a data instance, and let $\mathcal{D}$ be a probability distribution over $\{\pm 1\}^d$. Then, the precision error of any linear explanation $\boldsymbol{w}$ for $f(\boldsymbol{x})$ satisfies

$$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq \sqrt{\mathsf{F}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w})}$$

Proof. By sublinearity of the absolute loss function,

$$\begin{aligned}\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq\ & \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[|f(\boldsymbol{z}) - \boldsymbol{w}\cdot\boldsymbol{z}| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}\right] \\ &+ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[|\boldsymbol{w}\cdot\boldsymbol{z} - \boldsymbol{w}\cdot\boldsymbol{x}| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}\right] \\ &+ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[|\boldsymbol{w}\cdot\boldsymbol{x} - f(\boldsymbol{x})| \mid \boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}\right]\end{aligned}$$

Note that the second term in the above inequality vanishes. Since $\boldsymbol{w}$ satisfies the hyperplane condition $\boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x})$, the third term also disappears. Using the fact that the expectation in the first term is independent of the condition $\boldsymbol{w}\cdot\boldsymbol{z} = \boldsymbol{w}\cdot\boldsymbol{x}$, it follows from Jensen's Inequality that

$$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\left[|f(\boldsymbol{z}) - \boldsymbol{w}\cdot\boldsymbol{z}|\right] \leq \sqrt{\mathsf{F}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w})}$$

□

Importantly, (2) involves an *unconditional* expectation, which is approximable via sampling. In doing so, let $\{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$ be a sample set where each $\boldsymbol{z}_i$ is drawn independently at random according to $\mathcal{D}$, and its value $f(\boldsymbol{z}_i)$ is obtained through query access to f. The corresponding *empirical fidelity error* is given by

$$\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \tag{3}$$

Based on this objective function, (P3) is a variant of the well-studied problem known as *(L_0) sparse regression*, also referred to as *best subset selection*, which dates back at least to [Beale et al., 1967, Hocking and Leslie, 1967]. While the sparse regression problem is non-convex and NP-hard Natarajan [1995], the inspiring work by Bertsimas et al. [2016] has explored various Mixed Integer Programming (MIP) formulations. Using modern branch-and-cut solvers, the authors have empirically shown that probably optimal solutions for high-dimensional instances can often be found in a few hours. The next formulation is a variation of their parameter-free approach utilizing *Specially Ordered Sets* (SOS) [Bertsimas and Weismantel, 2005]:

$$\begin{aligned}\text{minimize}\quad & \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}\cdot\boldsymbol{z}_i - f(\boldsymbol{z}_i))^2 \\ \text{subject to}\quad & \boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x}) \\ & \boldsymbol{1}\cdot\boldsymbol{u} \leq k \\ & \|(w_j, 1-u_j)\|_0 \leq 1, \text{ for all } j\in[d] \\ & u_j \in \{0,1\} \text{ for all } j\in[d] \\ & w_j \in [-1,+1] \text{ for all } j\in[d]\end{aligned} \tag{MIP}$$

The last constraint is used to ensure that the set of k-sparse explanations is bounded. The following result shows that if the solver for (MIP) is supplied a number of samples m that is quadratic in k and logarithmic in d, then with high probability, the precision error of any returned solution is upper-bounded by the root of its empirical fidelity.

Theorem 2. Let $f : \{\pm1\}^d \to [-1,+1]$ be a regression model, $\boldsymbol{x} \in \{\pm1\}^d$ be a data instance, $\mathcal{D}$ be a probability distribution over $\{\pm1\}^d$, and $k \geq 1$ be a sparsity level. Then, for any k-sparse explanation $\boldsymbol{w} \in [-1,+1]^d$ for $f(\boldsymbol{x})$, any $\delta \in (0,1]$, and any $\varepsilon \in (0,1]$, if

$$m \geq \frac{1}{\varepsilon^4}\left(32\ln(2d) + 8\ln(\tfrac{2}{\delta})\right)(k+1)^2$$

then with probability at least $1-\delta$ over the choice of an i.i.d. sample set of size m,

$$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq \sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w})} + \varepsilon$$

Proof. Let $\mathcal{W}$ be the hypothesis class consisting of all vectors $\boldsymbol{w} \in [-1,+1]^d$ such that $\boldsymbol{w}\cdot\boldsymbol{x} = f(\boldsymbol{x})$ and $\|\boldsymbol{w}\|_0 \leq k$. Using here the fact that $\|\boldsymbol{w}\|_1 \leq \|\boldsymbol{w}\|_0$, we know that $\mathcal{W}$ is included in $\mathcal{B}_1(k)$. Additionally, let ℓ_f denote the loss function defined as $\ell_f(\boldsymbol{w},\boldsymbol{z}) = |\boldsymbol{w}\cdot\boldsymbol{z} - f(\boldsymbol{z})|$. By construction, $\ell_f(\boldsymbol{w},\boldsymbol{z})$ is 1-Lipschitz and upper-bounded by $k+1$ for all $\boldsymbol{z} \in \{0,1\}^d$. Therefore, by application of Theorem 26.15 in [Shalev-Shwartz and Ben-David, 2014] (see also Corollary 4 in [Kakade et al., 2008]), we have

$$\begin{aligned}\mathsf{F}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq\ & \widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}) \\ &+ 2(k+1)\sqrt{\frac{8\ln(2d) + 2\ln(\frac{2}{\delta})}{m}}\end{aligned}$$

By substituting the upper bound on m defined as above, and applying Lemma 1, the result follows. □

6 DEALING WITH NP-HARDNESS

In light of Theorem 2, we would like to find an *optimal* solution to (MIP), striving for the best possible empirical fidelity. However, since the sparse regression problem is NP-hard, we need to make some additional assumptions for achieving polynomial time efficiency, In this section, we focus on the *Restricted Isometry Property* (RIP) [Candes and Tao, 2005], a condition that is often recommended to overcome this computational challenge.

A matrix $\boldsymbol{Z} \in \mathbb{R}^{m\times d}$ is said to satisfy the RIP of order k with constant $\beta_k \in (0,1)$ if, for all vectors $\boldsymbol{w} \in \mathcal{B}_0(k)$, the following inequality holds:

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

This condition is equivalent to requiring that the Gram matrix of $\boldsymbol{Z}$, restricted to the columns in $\mathrm{supp}(\boldsymbol{w})$, is positive definite with its eigenvalues confined to the interval $[1-\beta_k, 1+\beta_k]$. Let $\mathcal{D}$ be a probability distribution over $\mathbb{R}^d$ such that for any $\boldsymbol{w} \in \mathbb{R}^d$ and any $\varepsilon \in (0,1)$, the following concentration inequality holds:

$$\mathbb{P}_{\boldsymbol{Z}\sim\mathcal{D}^m}\left[\left|\tfrac{1}{m}\|\boldsymbol{Z}\boldsymbol{w}\|_2^2 - \|\boldsymbol{w}\|_2^2\right| > \varepsilon\right] \leq 2e^{-\Omega(m)} \tag{4}$$

Algorithm 1: Iterative Hard Thresholding (IHT)

Input: query $(\boldsymbol{x}, f(\boldsymbol{x}))$, sparsity level k, data $(\boldsymbol{Z}, \boldsymbol{y})$

$\boldsymbol{w}_0 = \boldsymbol{0}$
for $t = 1, 2, \ldots$ **do**
  $\boldsymbol{v}_t = \boldsymbol{w}_{t-1} - \frac{1}{m}\boldsymbol{Z}^\intercal(\boldsymbol{Z}\boldsymbol{w}_{t-1} - \boldsymbol{y})$
  $\boldsymbol{w}_t = \Pi_{\mathcal{H}(\boldsymbol{x}, f(\boldsymbol{x})) \cap \mathcal{B}_0(k)}(\boldsymbol{v}_t)$

Algorithm 2: Projection onto k-Sparse Explanations

Input: query $(\boldsymbol{x}, y)$, sparsity level k, vector $\boldsymbol{w}$

$\boldsymbol{w}_H = \boldsymbol{w} \odot \boldsymbol{1}_{|\mathrm{supp}(\boldsymbol{x})}$ and $\boldsymbol{w}_B = \boldsymbol{w} \odot \boldsymbol{1}_{|[d] \setminus \mathrm{supp}(\boldsymbol{x})}$
$\boldsymbol{w}_B^* = \mathrm{HT}(\boldsymbol{w}_B, k)$
$\boldsymbol{w}_H^* = \mathrm{GSHP}(\boldsymbol{w}_H \odot \boldsymbol{x}, k, y) \odot \boldsymbol{x}$
return $\boldsymbol{w}_H^* + \boldsymbol{w}_B^*$

As shown in [Baraniuk et al., 2008], if $\mathcal{D}$ satisfies such a concentration inequality, then the RIP of order $k \leq \frac{d}{2}$ with constant β_k holds with probability at least $1 - 2e^{-\Omega(m)}$ for matrices $\boldsymbol{Z}$ drawn over $\mathcal{D}^m$, whenever $m = \Omega\big(\frac{k}{\beta_k^2} \ln \frac{d}{\beta_k}\big)$.

In the context of this study, we are interested in discrete distributions over $\{\pm 1\}^d$ satisfying (4). Under this assumption, our algorithm for computing k-sparse explanations of high fidelity is a variant of the *Iterative Hard Thresholding* (IHT) method [Blumensath and Davies, 2009, Garg and Khandekar, 2009, Jain et al., 2014]. Instead of projecting onto the ball $\mathcal{B}_0(k)$, it projects onto the intersection of this ball and the hyperplane $\mathcal{H}(\boldsymbol{x}, f(\boldsymbol{x}))$, ensuring that the solution serves as an explanation for $f(\boldsymbol{x})$.

As detailed in Algorithm 1, our version of IHT takes the following inputs: a data instance $\boldsymbol{x} \in \{\pm 1\}^d$ and its predicted value $f(\boldsymbol{x}) \in [-1, +1]$, along with a sparsity level $k \geq 1$. Additionally, the algorithm requires a sample set $\{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$, which is compactly represented as a pair $(\boldsymbol{Z}, \boldsymbol{y})$, where $\boldsymbol{Z} \in \{\pm 1\}^{m \times d}$ is the matrix of samples $\boldsymbol{z}_i$, and $\boldsymbol{y} \in [-1, +1]^m$ is the vector of the corresponding labels $f(\boldsymbol{z}_i)$. The algorithm performs gradient descent (with a step size of 1), followed by a projection onto the set of k-sparse explanations. The following result ensures that each iteration of the algorithm operates in low polynomial time.

Lemma 2. For a model f, an instance $\boldsymbol{x}$, a sparsity level k, and a vector $\boldsymbol{w}$, the projection of $\boldsymbol{w}$ onto $\mathcal{H}(\boldsymbol{x}, f(\boldsymbol{x})) \cap \mathcal{B}_0(k)$ can be computed in $\mathcal{O}(d \log_2(d) + k^2)$ time.

Proof. As outlined in Algorithm 2, the idea is to split $\boldsymbol{w}$ into two components: one that depends on the hyperplane constraint and another that does not. Specifically, let $\boldsymbol{w} = \boldsymbol{w}_H + \boldsymbol{w}_B$, where $\boldsymbol{w}_H$ is the projection of $\boldsymbol{w}$ onto the support set S of $\boldsymbol{x}$, and $\boldsymbol{w}_B$ is the projection of $\boldsymbol{w}$ onto the complement of S. Since all indices in $\boldsymbol{w}_B$ are free variables in the equation $\boldsymbol{w} \cdot \boldsymbol{x} = y$, we can directly project $\boldsymbol{w}_B$ onto $\mathcal{B}_0(k)$. The solution $\boldsymbol{w}_B^*$ can be obtained in $\mathcal{O}(d \log_2(d))$ time using the *Hard Thresholding* (HT) operator, which sets all but the largest (in magnitude) elements of $\boldsymbol{w}_B$ to zero.

Now, let $\boldsymbol{u}_H = \boldsymbol{w}_H \odot \boldsymbol{x}$ and let $y = f(\boldsymbol{x})$. In addition, let $\mathcal{W}$ and $\mathcal{U}$ denote the intersections of the ball $\mathcal{B}_0(k)$ with the hyperplanes $\mathcal{H}(\boldsymbol{x}, y)$ and $\mathcal{H}(\boldsymbol{1}_S, y)$, respectively. Since $\boldsymbol{u}_H \cdot \boldsymbol{1}_S = y$ if and only if $\boldsymbol{w}_H \cdot \boldsymbol{x} = y$, it follows that $\boldsymbol{u}' \in \mathcal{U}$ if and only if $\boldsymbol{w}' = (\boldsymbol{u} \odot \boldsymbol{x}) \in \mathcal{W}$. This, together with the fact that $\|\boldsymbol{u}' - \boldsymbol{u}_H\|_2 = \|\boldsymbol{w}' - \boldsymbol{w}_H\|_2$, implies that $\Pi_{\mathcal{W}}(\boldsymbol{w}_H) = (\Pi_{\mathcal{U}}(\boldsymbol{u}_H)) \odot \boldsymbol{x}$. Let $\boldsymbol{v}_H$ be the projection of $\boldsymbol{u}_H$ onto $\mathcal{U}$. By setting $\boldsymbol{w}_H^* = (\boldsymbol{v}_H \odot \boldsymbol{x})$, the projection of $\boldsymbol{w}$ onto $\mathcal{W}$ is therefore $\boldsymbol{w}^* = \boldsymbol{w}_H^* + \boldsymbol{w}_B^*$.

Finally, since $\mathcal{H}(\boldsymbol{1}_S, \lambda)$ is a diagonal hyperplane, the runtime complexity for deriving $\boldsymbol{w}_H^*$ follows from the fact that $\boldsymbol{v}_H$ can be obtained in $\mathcal{O}(d \log_2(d) + k^2)$ time using the *Greedy Selector and Hyperplane Projector* (GSHP) operation [Kyrillidis et al., 2013]. □

With this lemma in hand, the main result of this section can be formally stated in the following theorem.

Theorem 3. Let $f : \{\pm 1\}^d \to [-1, +1]$ be a regression model, $\boldsymbol{x} \in \{\pm 1\}^d$ be a data instance, $k \in [1, \frac{d}{6}]$ be a sparsity level, and $\mathcal{D}$ be a probability distribution over $\{\pm 1\}^d$ satisfying the concentration inequality (4). Suppose that the IHT algorithm is run on a sample set $\{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$ drawn from $\mathcal{D}$ and labeled by f such that $m = \Omega\big(\frac{k}{\alpha^2} \ln \frac{d}{\alpha}\big)$ with $\alpha < 1/(32\sqrt{3})$. Then, for any k-sparse explanation $\boldsymbol{w}$ for $f(\boldsymbol{x})$, after

$$t \geq \log_2 \left\lceil \frac{\|\boldsymbol{w}\|_2}{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w})} \right\rceil$$

iterations, the returned vector $\boldsymbol{w}_t$ is a k-sparse explanation for $f(\boldsymbol{x})$ satisfying, with probability at least $1 - 2e^{-\Omega(m)}$,

$$\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}_t)} \leq 7\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w})}$$

Proof. Let $\boldsymbol{Z} \in \{\pm 1\}^{m \times d}$ be the matrix of samples $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ and let $\boldsymbol{y} \in [-1, +1]^m$ be the vector of corresponding values $(f(\boldsymbol{z}_1), \ldots, f(\boldsymbol{z}_m))$. By applying Lemma 5.1 from [Baraniuk et al., 2008] and using the bound on m, we know that with a probability of at least $1 - 2e^{-\Omega(m)}$, the matrix $\boldsymbol{Z}$ satisfies the RIP of order $3k$ with a constant $\beta_{3k} < \frac{1}{32}$. By integrating this result with Theorem 5 from [Blumensath and Davies, 2009], we can conclude that at iteration t, defined as above, the solution $\boldsymbol{w}_t$ computed by IHT satisfies, with probability at least $1 - 2e^{-\Omega(m)}$,

$$\|\boldsymbol{w}_t - \boldsymbol{w}\|_2 \leq 6\|\boldsymbol{e}\|_2, \text{ where } \boldsymbol{e} = \tfrac{1}{m}\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{y}$$

Leveraging this result, and applying the triangle inequality

along with the fact that $\frac{1}{m}\|\boldsymbol{Z}\boldsymbol{u}\|_2 \leq \|\boldsymbol{u}\|_2$, we obtain

$$\begin{aligned}\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}_t)} &= \tfrac{1}{m}\|\boldsymbol{Z}\boldsymbol{w}_t + \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{Z}\boldsymbol{w} - \boldsymbol{y}\|_2 \\ &\leq \tfrac{1}{m}\|\boldsymbol{Z}(\boldsymbol{w}_t - \boldsymbol{w})\|_2 + \tfrac{1}{m}\|\boldsymbol{Z}\boldsymbol{w} - \boldsymbol{y}\|_2 \\ &\leq \|\boldsymbol{w}_t - \boldsymbol{w}\|_2 + \|\boldsymbol{e}\|_2 \\ &\leq 7\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w})}\end{aligned}$$

□

As shown in [Achlioptas, 2001], the uniform distribution $\mathcal{U}$ over $\{\pm 1\}^d$ satisfies the concentration inequality (4). Thus, by combining Theorems 2 and 3, we know that 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

$$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}_t) \leq 7\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}^*)} + o(1),$$

where $\boldsymbol{w}^*$ is the optimal solution to (MIP). Additionally, by integrating Lemma 2 and the fact that $\|\boldsymbol{w}^*\|_\infty \leq 1$, we can conclude that the solution $\boldsymbol{w}_t$ can be computed in polynomial time with respect to d, k, and $\log_2\lceil 1/\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}^*)\rceil$.

At first glance, this result may seem surprising because, as indicated in Theorem 1, finding k-sparse linear explanations with minimal precision is $\mathsf{NP}^{\mathsf{PP}}$-hard. However, it is important to keep in mind that the fidelity measure does not always provide a tight upper bound on the precision measure. In fact, since $\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}^*)$ here assesses the capability of $\boldsymbol{w}^*$ to fit the regression model f over the uniform distribution, it can be quite large.

7 EXPERIMENTS

In order to validate the effectiveness of our methods, we have considered various explanation tasks for regression models. The code was written using the `Python` language. Our experiments have been conducted on a Quad-core Intel XEON X5550 with 32GB of memory.

7.1 EXPERIMENTAL SETUP

We conducted experiments using 18 tabular datasets, sourced from the standard repository, `OpenML`.[2] All datasets focus on regression tasks and include both numerical and categorical attributes. To convert these raw attributes into interpretable binary features, we applied a standard K-bins discretization method, creating 4 bins for each attribute. For our experimental purposes, the 18 datasets were divided into two groups: 12 medium-dimensional benchmarks with an average of 415 binary features, and 6 low-dimensional benchmarks with an average of 20 binary features.

For each benchmark, an explanation task is defined by a tuple $(f, \boldsymbol{x}, \sigma, k)$, where f is a black-box regressor implemented using a neural network. In our experiments, we utilized the `Scikit-Learn` implementation of the multi-layer perceptron regressor with default parameters. As usual, we trained f on the training set of the benchmark and evaluated its accuracy on the test set. Each data instance $\boldsymbol{x}$ that we aimed to explain was randomly selected from the test set using a uniform distribution. Since the performance of state-of-the-art model-agnostic explainers is evaluated according to neighborhood distributions around $\boldsymbol{x}$, we employed the following parameterized distribution:

$$\mathcal{D}_{\boldsymbol{x},\sigma}(\boldsymbol{z}) = \frac{1}{Z_\sigma} e^{-\sigma\|\boldsymbol{x}-\boldsymbol{z}\|_1} \quad \text{where} \quad Z_\sigma = \sum_{j=0}^{d} \binom{d}{j} e^{-\sigma j}$$

Here, $\sigma \geq 0$ serves as a spread parameter. Note that $\mathcal{D}_{\boldsymbol{x},0}$ corresponds to the uniform distribution. We also considered $k \in \{1, \ldots, 10\}$ to explore different levels of sparsity.

The performance of explainers for each explanation task was measured using the root mean squared error $(\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w}))^{\frac{1}{2}}$ of the generated explanation $\boldsymbol{w}$. This metric was calculated using $m = 1000$ labeled samples $(\boldsymbol{z}_i, f(\boldsymbol{z}_i))$, where each $\boldsymbol{z}_i$ was generated according to the distribution $\mathcal{D}_{\boldsymbol{x},\sigma}$. For low-dimensional benchmarks, we also calculated the precision error $\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w})$ of the generated explanation $\boldsymbol{w}$ by enumerating all data instances $\boldsymbol{z} \in \{\pm 1\}^d$. Both metrics were averaged over 20 random instances $\boldsymbol{x}$.

To implement the MIP approach specified in the formulation (MIP), we used the `Gurobi` solver (version 11.0), running on a single thread with a timeout of 60 seconds. Our MIP and IHT approaches were compared with three methods. The first, referred to as CVX, is the convex relaxation of (MIP), obtained by replacing the constraint $\|\boldsymbol{w}\|_0 \leq k$ with $\|\boldsymbol{w}\|_1 \leq k$. The last two methods are the state-of-the-art LIME [Ribeiro et al., 2016] and MAPLE [Plumb et al., 2018], both implemented with default parameters.

7.2 EXPERIMENTAL RESULTS

An overview of our experimental results on 18 benchmarks, specifically for $\sigma = 1$ and $k = 7$, is presented in Table 1. The first six rows report results for the low-dimensional benchmarks, while the last twelve rows cover the medium-dimensional benchmarks. The first two columns of the table include the name of each dataset along with its corresponding `OpenML` identifier. The last five columns display the average root mean squared errors for the explanations generated by different competitors. Entries highlighted in blue indicate that the sparsity of all inferred explanations is at most k, while entries highlighted in black indicate that the sparsity of the explanations significantly exceeds k.

[2] Some statistics on the datasets used in our experiments can be found in Table 3 of Appendix B.

Benchmark		$\sqrt{\widehat{\mathsf{F}}_{f,\boldsymbol{x},m}(\boldsymbol{w})}$				
Name	ID	CVX	IHT	LIME	MAPLE	MIP
Airfoil Self Noise	44957	0.040 (±0.01)	0.055 (±0.02)	0.321 (±0.04)	0.218 (±0.02)	**0.049 (±0.01)**
Auto MPG	42372	0.031 (±0.00)	0.069 (±0.02)	0.338 (±0.07)	0.122 (±0.05)	**0.039 (±0.01)**
Bike Sharing	44142	0.040 (±0.00)	0.080 (±0.01)	0.183 (±0.05)	0.121 (±0.03)	**0.048 (±0.01)**
Liver Disorders	8	0.059 (±0.02)	0.091 (±0.02)	0.209 (±0.04)	0.147 (±0.08)	**0.068 (±0.02)**
Machine CPU	230	0.039 (±0.01)	0.128 (±0.02)	0.312 (±0.08)	0.190 (±0.06)	**0.055 (±0.01)**
Medical Charges	44146	0.040 (±0.00)	**0.049 (±0.01)**	0.408 (±0.01)	0.204 (±0.01)	**0.049 (±0.00)**
Ailerons	44137	0.050 (±0.01)	0.201 (±0.02)	0.647 (±0.05)	0.113 (±0.02)	**0.085 (±0.02)**
Auto Imports	9	0.067 (±0.01)	0.232 (±0.03)	0.528 (±0.06)	0.148 (±0.04)	**0.107 (±0.01)**
DNA Methylation	46139	0.121 (±0.02)	**0.192 (±0.04)**	0.582 (±0.08)	0.168 (±0.04)	**0.191 (±0.01)**
Geographical OM	44965	0.148 (±0.02)	0.259 (±0.04)	0.662 (±0.07)	0.174 (±0.01)	**0.202 (±0.02)**
Moneyball	41021	0.039 (±0.00)	0.192 (±0.02)	0.483 (±0.05)	0.120 (±0.03)	**0.071 (±0.01)**
NCI 60 Thioguanine	46132	0.062 (±0.01)	0.235 (±0.07)	0.534 (±0.10)	0.108 (±0.02)	**0.132 (±0.06)**
Online News	42724	0.010 (±0.00)	0.051 (±0.00)	0.069 (±0.01)	0.046 (±0.01)	**0.028 (±0.01)**
Pollution	542	0.045 (±0.01)	0.171 (±0.05)	0.478 (±0.06)	0.133 (±0.06)	**0.082 (±0.02)**
RTE Consumption	46337	0.033 (±0.00)	0.102 (±0.01)	0.273 (±0.10)	0.114 (±0.04)	**0.057 (±0.01)**
Student Performance	42352	0.074 (±0.01)	0.143 (±0.02)	0.454 (±0.03)	0.169 (±0.04)	**0.105 (±0.01)**
Wave Energy	44975	0.017 (±0.00)	**0.080 (±0.03)**	0.301 (±0.03)	0.128 (±0.02)	0.091 (±0.01)
Wisconsin	191	0.075 (±0.01)	0.135 (±0.02)	0.275 (±0.08)	0.201 (±0.05)	**0.111 (±0.02)**

Table 1: Experimental results on 6 low-dimensional benchmarks (upper rows) and 12 medium-dimensional benchmarks (lower rows), using $\sigma = 1$, and $k = 7$. Entries highlighted in blue indicate that all generated explanations were k-sparse.

From these results, we can confidently conclude that both the IHT and MIP approaches outperform LIME across all benchmarks. As CVX operates within the convex hull of k-sparse explanations (with $\|\boldsymbol{w}\|_\infty \leq 1$), it serves as a lower bound for the fidelity of solutions to (MIP). However, because CVX tends to produce dense solutions, it cannot be effectively considered as an explainer. Additionally, we can observe that for all medium-dimensional benchmarks, the explanations generated by MAPLE are dense. Furthermore, on low-dimensional benchmarks, MAPLE consistently performs worse than both IHT and MIP.

Table 2 presents the average precision errors of IHT, LIME, and MIP across six low-dimensional benchmarks. By comparing these results with the average root mean square errors shown in Table 1, we can see that empirical fidelity serves as a good indicator of an explainer's performance regarding precision errors. Notably, both IHT and MIP outperform LIME in terms of precision.

In Figure 2, we present the performance of IHT, LIME, and MIP across varying levels of sparsity k, ranging from 1 to 10, on three benchmarks: *DNA Methylation* (which affects cancer drug response), *Student Performance*, and *Wave Energy*. The bar plots reveal that the performance of both IHT and MIP remains stable or even improves as k increases, while LIME exhibits significantly less stability.

Additionally, in Figure 3, we report the performance of the three explainers as the spread σ increases from 0.1 to 1.0, using the same benchmarks. In contrast to LIME, both IHT and MIP show robustness to variations in the distribution.

Benchmark	$\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w})$		
Name	IHT	LIME	MIP
Airfoil S. N.	0.045 (±0.03)	0.092 (±0.07)	0.042 (±0.01)
Auto MPG	0.028 (±0.02)	0.063 (±0.02)	0.019 (±0.01)
Bike Sharing	0.067 (±0.03)	0.101 (±0.08)	0.041 (±0.01)
Liver Disorders	0.024 (±0.03)	0.071 (±0.05)	0.010 (±0.02)
Machine CPU	0.088 (±0.02)	0.124 (±0.09)	0.035 (±0.01)
Medical Charges	0.012 (±0.01)	0.228 (±0.04)	0.012 (±0.01)

Table 2: Average precisions of IHT, LIME, and MIP across the 6 low-dimensional benchmarks.

The runtimes of the explainers are outlined in Appendix B. In summary, the CVX and LIME methods are the fastest, each taking only a few milliseconds per benchmark. The IHT and MAPPLE methods have comparable speeds, generally requiring a few seconds per benchmark. For the MIP approach, the `Gurobi` solver can find an optimal solution within a few seconds for low-dimensional benchmarks. However, it constantly reaches the one-minute timeout for medium-dimensional benchmarks. In these cases, we have found that `Gurobi` can identify near-optimal solutions in just a few seconds, but verifying their optimality through lower bounds may take several minutes.

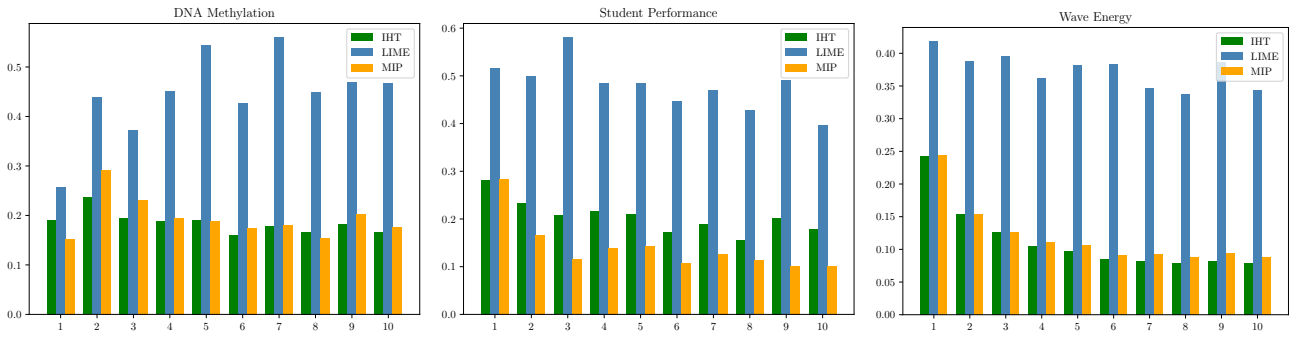

Figure 2: Comparison of root mean squared errors (y-axis) with increasing sparsity level (x-axis).

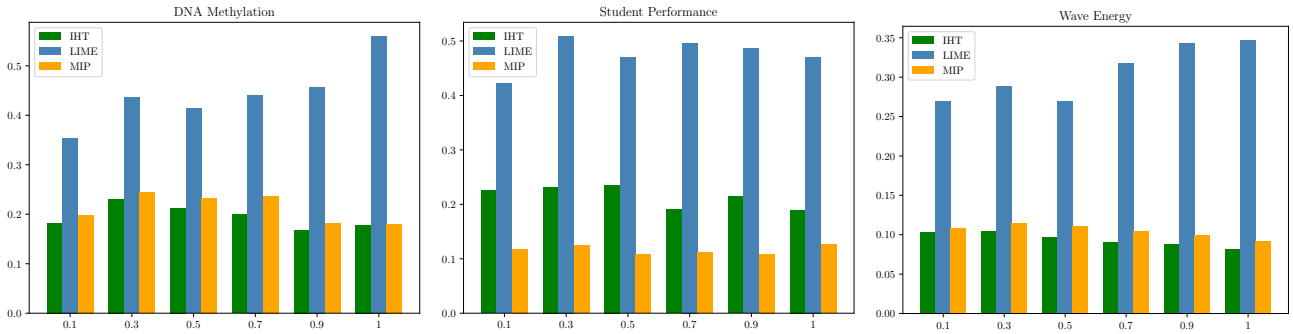

Figure 3: Comparison of root mean squared errors (y-axis) with increasing spread (x-axis).

8 CONCLUSIONS

In this paper, we have demonstrated that deriving sparse and precise explanations for regression models is $\mathsf{NP}^{\mathsf{PP}}$-hard. To tackle this computational challenge, we established that the precision of these explanations is upper-bounded by their fidelity. We can address this surrogate objective using Mixed Integer Programming, and under certain assumptions about the underlying distribution, we can achieve polynomial time efficiency through Iterative Hard Thresholding. Our comparative experiments on real-world regression tasks support these theoretical findings.

Though this study focused on minimizing the precision $\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w})$ of linear explanations while maintaining a desired level of sparsity $\|\boldsymbol{w}\|_0 \leq k$, a promising direction for future research is to explore the reverse problem: minimizing the sparsity of linear explanations $\|\boldsymbol{w}\|_0$ while ensuring that the desired precision $\mathsf{P}_{f,\boldsymbol{x},\mathcal{D}}(\boldsymbol{w}) \leq \epsilon$ is maintained. This latter problem is also challenging, as verifying such a probabilistic constraint is PP-hard.

Acknowledgements. Many thanks to the reviewers for their comments and suggestions. This work has benefited from the support of the AI Chair EXPEKCTATION (ANR-19- CHIA-0005-01) of the French National Research Agency. It was also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

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

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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 110–119, 2021.

Marcelo Arenas, Pablo Barceló, Miguel A. Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees,. In *Advances in Neural Information Processing Systems 35: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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.

E. M. L. Beale, M. G. Kendall, and D. W. Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366, 1967.

Dimitris Bertsimas and Robert Weismantel. *Optimization Over Integers*. Dynamic Ideas, 2005.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Guy Blanc, Jane Lange, and Li-Yang Tan. Provably efficient, succinct, and precise explanations. In *Advances in Neural Information Processing Systems 34: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 6129–6141, 2021.

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Louenas Bounia and Frederic Koriche. Approximating probabilistic explanations via supermodular minimization. In *Proceedings of the 39th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 216–225, 2023.

Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Martin Cooper and João Marques-Silva. Tractability of explaining classifier decisions. *Artificial Intelligence*, 316:103841, 2023.

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pages 712–720, 2020.

Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, page 337–344, 2009.

R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

Alexey Ignatiev. Towards trustable explainable AI. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5154–5158, 2020.

Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *International Journal of Approximate Reasoning*, 159:108939, 2023.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems 27: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, page 685–693, 2014.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21, Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 793–800, 2008.

Frederic Koriche, Jean-Marie Lagniez, Stefan Mengel, and Chi Tran. Learning model agnostic explanations via constraint programming. In *Machine Learning and Knowledge Discovery in Databases. Research Track (ECML/PKDD)*, pages 437–453, 2024.

Anastasios Kyrillidis, Stephen Becker, Volkan Cevher, and Christoph Koch. Sparse projections onto the simplex. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 235–243, 2013.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 59–67, 2019.

Jeffrey Li, Vaishnavh Nagarajan, Gregory Plumb, and Ameet Talwalkar. A learning theoretic perspective on local explainability. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Joao Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. *Proceedings of the 36th Annual AAAI Conference on Artificial Intelligence*, pages 12342–12350, 2022.

Georges A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

Tim Miller, Robert R. Hoffman, Ofra Amir, and Andreas Holzinger. Special issue on explainable artificial intelligence (XAI). *Artif. Intell.*, 307:103705, 2022.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. leanpub.com, 2nd edition, 2022.

Balas K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems 31. Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2520–2529, 2018.

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.

Thomas L. Saaty and Müjgan Sagir Özdemir. Why the magic number seven plus or minus two. *Mathematical and Computer Modelling*, 38(3):233–244, 2003.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Bernardo Subercaseaux, Marcelo Arenas, and Kuldeep S Meel. Probabilistic explanations for linear models. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, pages 20655–20662, 2025.

Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70:351–387, 2021.

Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. BayLIME: Bayesian local interpretable model-agnostic explanations. In *Proceedings of the 37th International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 887–896, 2021.

Probabilistic Explanations for Regression Models (Supplementary Material)

Frederic Koriche[1] **Jean-Marie Lagniez**[1] **Chi Tran**[1]

[1]Univ. Artois, CNRS, Centre de Recherche en Informatique de Lens (CRIL), France

A ADDITIONAL THEORETICAL BACKGROUND

Our main results in Sections 5 and 6 are based on Theorem 26.15 in [Shalev-Shwartz and Ben-David, 2014] and Theorem 5 in [Blumensath and Davies, 2009], which are presented below.

Given a space of data instances $\mathcal{X} \subseteq \mathbb{R}^d$, a space of labels $\mathcal{Y} \subseteq \mathbb{R}$, and a hypothesis class of linear functions $\mathcal{H} \subseteq \mathbb{R}^d$, let $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function of the form

$$\ell(\boldsymbol{w}, x, y) = \phi(\boldsymbol{w} \cdot \boldsymbol{x}, y) \tag{5}$$

where $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is ρ-Lipschitz in its first argument. In other words, for every $y \in \mathcal{Y}$, the scalar function $a \mapsto \phi(a, y)$ is ρ-Lipschitz. As a notable example, the absolute loss function given by $\ell(\boldsymbol{w}, x, y) = |\boldsymbol{w} \cdot \boldsymbol{x} - f(\boldsymbol{x})|$ can be written as in Equation 5 using $\phi(a, y) = |a - y|$, which is 1-Lipschitz for all $y \in \mathbb{R}$.

Theorem 4 ([Shalev-Shwartz and Ben-David, 2014])**.** Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ such that $\|\boldsymbol{x}\|_\infty \leq r$ with probability 1. Additionally, let $\mathcal{H} = \mathcal{B}_0(b)$ and let $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function of the form given in Equation 5, such that ϕ is ρ-Lipschitz in it first argument, and such that $\max_{a \in [-br, +br]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample set $\boldsymbol{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$,

$$\forall \boldsymbol{w} \in \mathcal{H}, \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(\boldsymbol{w}, \boldsymbol{x}, y)] \leq \frac{1}{m} \sum_{i=1}^{m} \ell(\boldsymbol{w}, \boldsymbol{x}_i, y_i) + 2\rho br \sqrt{\frac{2 \log_2(2d)}{m}} + c\sqrt{\frac{2 \ln(2/\delta)}{m}}.$$

Recall that a matrix $\boldsymbol{X} \in \mathcal{X}^m$ satisfies the RIP of order s with constant $\beta_s \in (0, 1)$ if, for any vector $\boldsymbol{w} \in \mathcal{B}_0(s)$, the following inequality holds:

$$(1 - \beta_s)\|\boldsymbol{w}\|_2^2 \leq \tfrac{1}{m}\|\boldsymbol{X}\boldsymbol{w}\|_2^2 \leq (1 + \beta_s)\|\boldsymbol{w}\|_2^2$$

Theorem 5 ([Blumensath and Davies, 2009])**.** Consider a noisy observation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{e}$ where $\boldsymbol{w} \in \mathcal{B}_0(k)$. If $\boldsymbol{X}$ has the RIP of order $s = 3k$ with constant $\beta_s < 1/\sqrt{32}$, then after at most

$$t = \left\lceil \log_2 \left(\frac{\|\boldsymbol{w}\|_2}{\|\boldsymbol{e}\|_2} \right) \right\rceil$$

iterations, the solution $\boldsymbol{w}_t$ returned by the IHT algorithm estimates $\boldsymbol{w}$ with accuracy

$$\|\boldsymbol{w}_t - \boldsymbol{w}\|_2 \leq 6\|\boldsymbol{e}\|_2.$$

B ADDITIONAL EXPERIMENTAL RESULTS

Table 3 presents statistics for the 18 benchmarks used in our experiments. The first two columns list the names of the datasets along with their corresponding OPENML identifiers. The next four columns provide information on the number of

categorical attributes, the number of numeric attributes, the count of binarized literals, and the total number of instances. Finally, the last column displays the accuracy of the regression model f, measured as the mean squared error and obtained through 10-fold cross-validation.

Benchmark		Qualities				Accuracy
Name	ID	#CAT	#NUM	#BIN	#INST	(MSE)
Airfoil Self Noise	44957	0	5	20	1503	0.104
Auto MPG	42372	0	5	20	392	0.054
Bike Sharing (Demand)	44142	0	6	24	17379	0.090
Liver Disorders	8	0	5	20	345	0.104
Machine CPU	230	0	6	24	209	0.029
Medical Charges	44146	0	3	12	163065	0.106
Ailerons	44137	0	33	132	13750	0.037
Auto Imports	9	11	14	120	205	0.027
DNA Methylation	46139	0	808	3232	475	0.028
Geographical OM	44965	0	116	464	1059	0.019
Moneyball	41021	9	5	96	1232	0.029
NCI 60 Thioguanine	46132	0	48	192	60	0.021
Online News (Popularity)	42724	0	59	208	39644	0.004
Pollution	542	0	15	60	60	0.025
RTE Consumption	46337	0	15	56	105168	0.036
Student Performance	42352	0	32	103	395	0.020
Wave Energy	44975	0	48	192	72000	0.024
Wisconsin	191	0	32	128	194	0.027

Table 3: Some statistics about the 18 benchmarks.

Table 4 provides the runtimes in seconds for all explainers. As indicated in the paper, the MIP approach, which uses the `Gurobi` solver, is capable of finding an optimal solution within a few seconds for low-dimensional benchmarks. However, it experiences a one-minute timeout (highlighted in red) when applied to medium-dimensional benchmarks.

Figures 4 and 5 show bar plots illustrating the increasing sparsity for all datasets, while Figures 6 and 7 present bar plots depicting the increasing spread for all datasets. Lastly, the plots in Figure 8 illustrate the evolution of the solution maintained by the `Gurobi` solver as the time budget increases. It is evident that a near-optimal solution is typically found within just a few seconds, with the majority of the time budget dedicated to certifying its optimality through lower bounds.

Benchmark		Time (s)				
Name	ID	CVX	IHT	LIME	MAPLE	MIP
Airfoil Self Noise	44957	0.005	0.238	0.015	0.177	1.178
Auto MPG	42372	0.005	0.241	0.020	0.185	1.125
Bike Sharing	44142	0.004	0.292	0.015	0.210	6.192
Liver Disorders	8	0.004	0.239	0.015	0.175	1.295
Machine CPU	230	0.004	0.303	0.020	0.212	2.920
Medical Charges	44146	0.004	0.145	0.011	0.125	0.601
Ailerons	44137	0.018	2.110	0.056	1.412	60.00
Auto Imports	9	0.015	1.494	0.043	0.985	60.00
DNA Methylation	46139	0.315	8.298	0.249	5.292	60.01
Geographical OM	44965	0.270	7.055	0.171	4.142	60.02
Moneyball	41021	0.009	0.817	0.038	0.637	60.01
NCI 60 Thioguanine	46132	0.028	2.790	0.084	1.853	60.01
Online News	42724	0.026	3.150	0.093	1.292	60.00
Pollution	542	0.009	0.993	0.030	0.623	60.00
RTE Consumption	46337	0.008	0.814	0.024	0.581	60.01
Student Performance	42352	0.020	1.817	0.044	1.208	60.00
Wave Energy	44975	0.028	2.395	0.066	1.905	60.00
Wisconsin	191	0.020	1.628	0.074	0.967	60.01

Table 4: Average runtimes for all explainers across the 18 benchmarks.

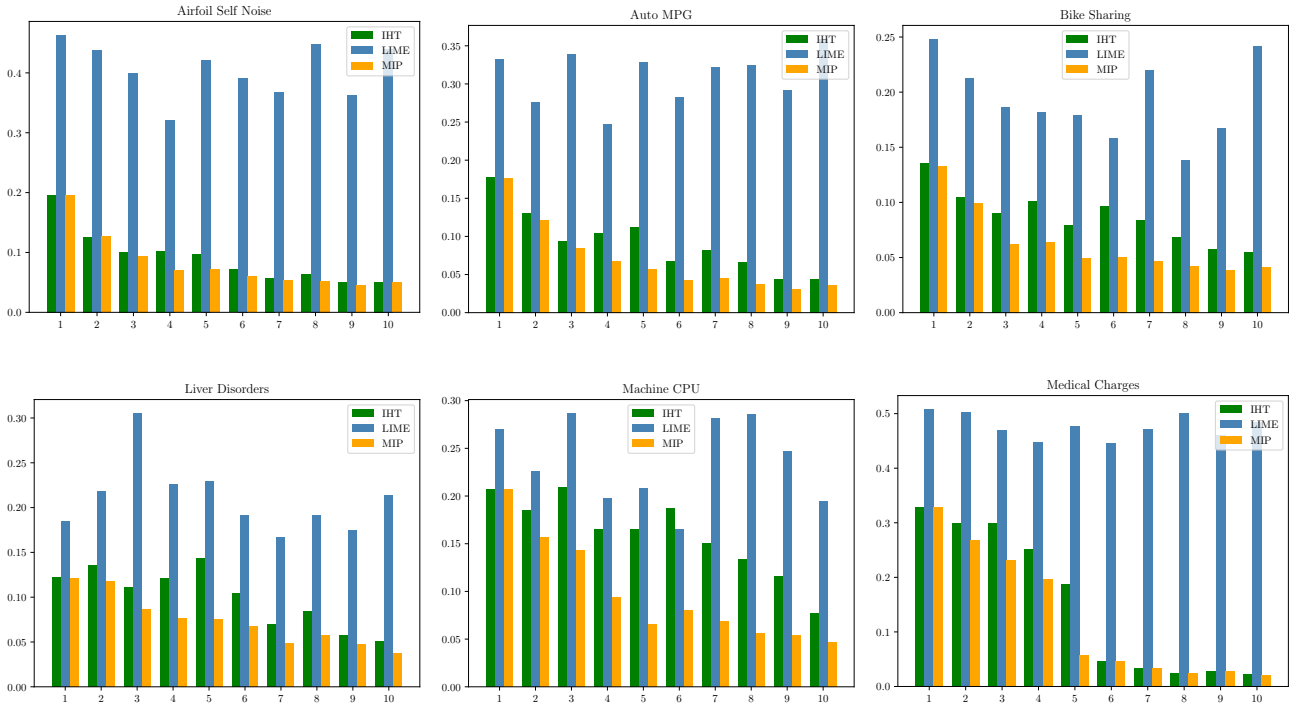

Figure 4: Comparison of root mean squared errors (y-axis) with increasing sparsity level (x-axis): Low-dimensional benchmarks.

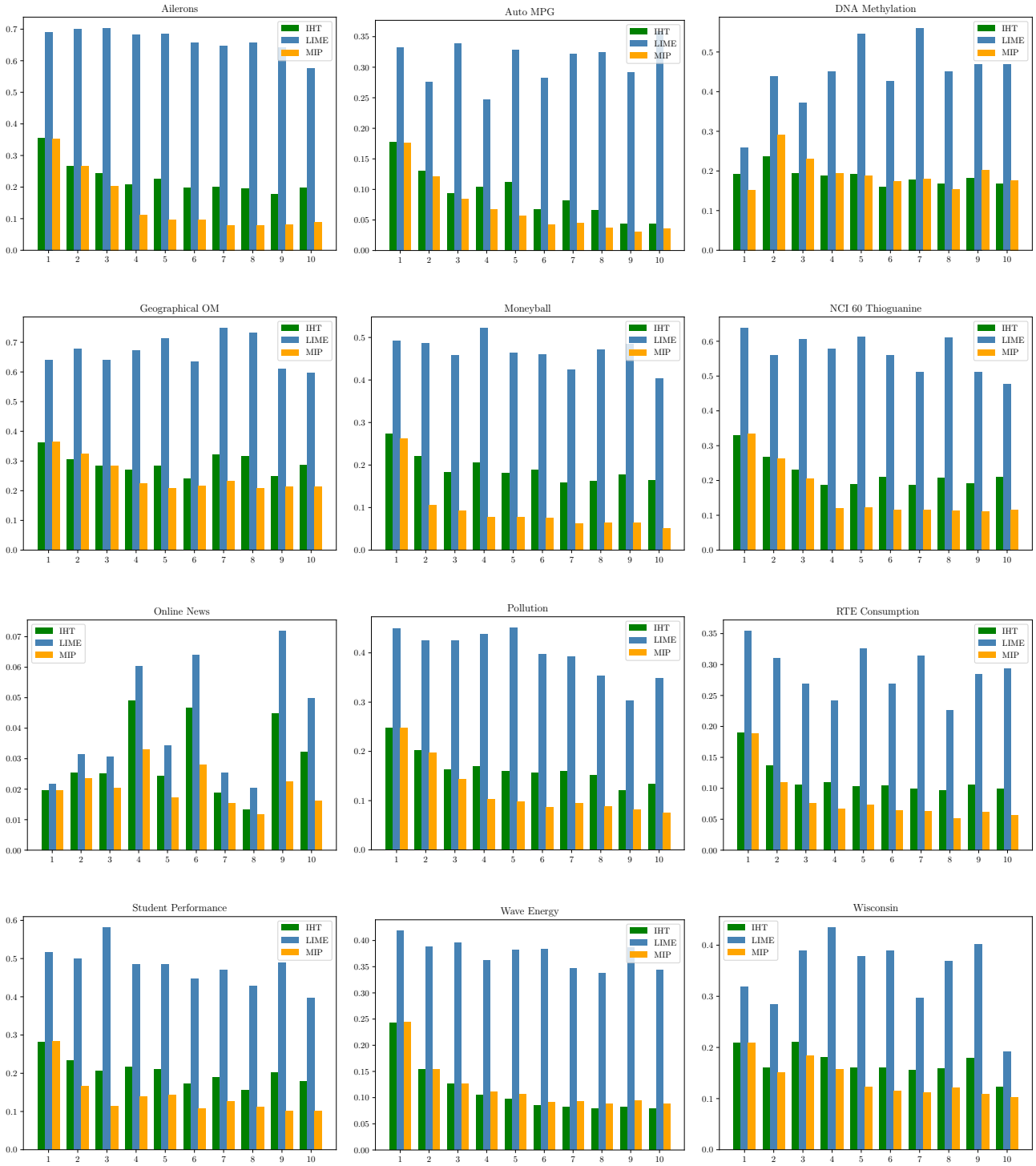

Figure 5: Comparison of root mean squared errors (y-axis) with increasing sparsity level (x-axis): Medium-dimensional benchmarks.

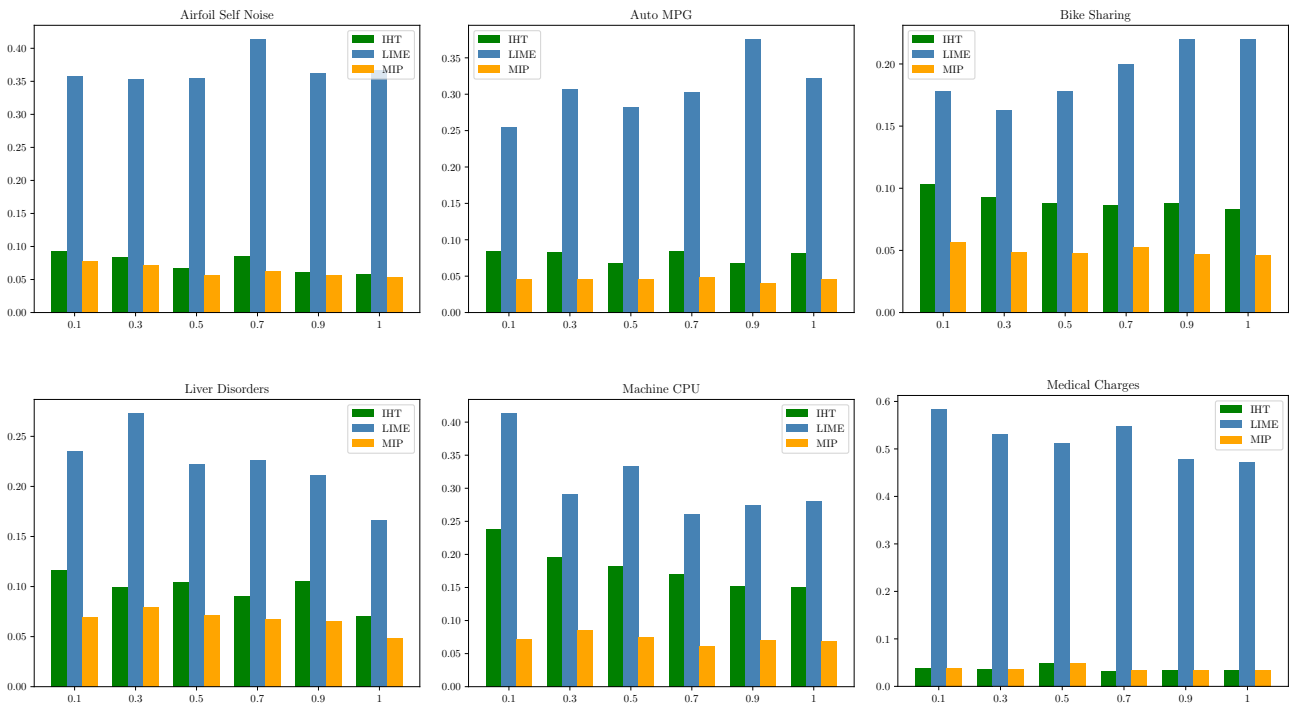

Figure 6: Comparison of root mean squared errors (y-axis) with increasing spread (x-axis): Low-dimensional benchmarks.

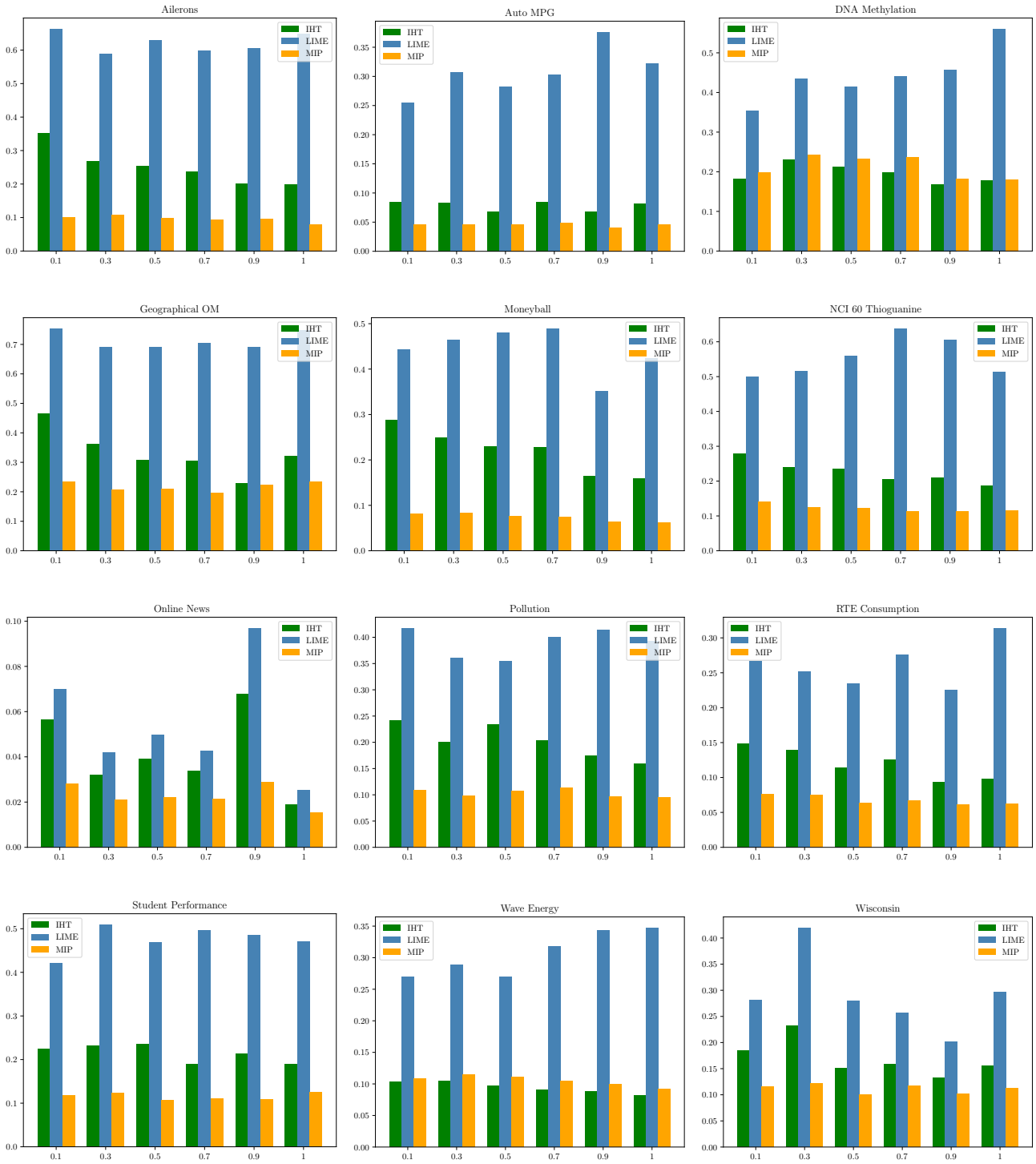

Figure 7: Comparison of root mean squared errors (y-axis) with increasing spread (x-axis): Medium-dimensional benchmarks.

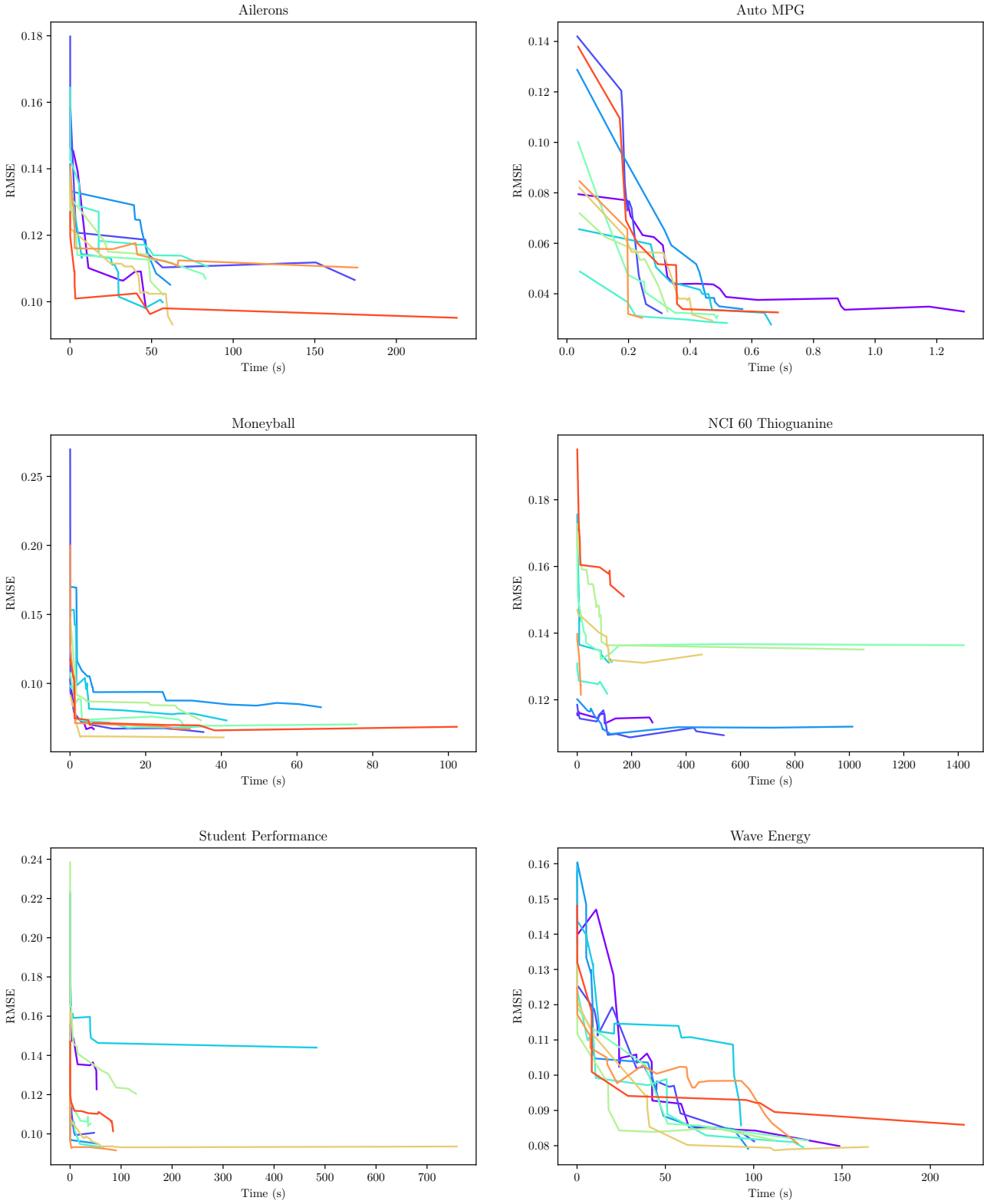

Figure 8: Evolution of the explanations computed by `Gurobi` for 10 data instances $\boldsymbol{x}$, using $k = 7$ and $\sigma = 1$.