

---

# THE PRICE OF INTERPRETABILITY

---

A PREPRINT

**Dimitris Bersimas**

Sloan School of Management  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
dbertsim@mit.edu

**Arthur Delarue**

Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
adelarue@mit.edu

**Patrick Jaillet**

Dep. of Electrical Engineering and Computer Science  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
jaillet@mit.edu

**Sebastien Martin**

Operations Research Center  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
92sebastien@gmail.com

July 9, 2019

## ABSTRACT

When quantitative models are used to support decision-making on complex and important topics, understanding a model’s “reasoning” can increase trust in its predictions, expose hidden biases, or reduce vulnerability to adversarial attacks. However, the concept of interpretability remains loosely defined and application-specific. In this paper, we introduce a mathematical framework in which machine learning models are constructed in a sequence of *interpretable steps*. We show that for a variety of models, a natural choice of interpretable steps recovers standard interpretability proxies (e.g., sparsity in linear models). We then generalize these proxies to yield a parametrized family of consistent measures of model interpretability. This formal definition allows us to quantify the “price” of interpretability, i.e., the tradeoff with predictive accuracy. We demonstrate practical algorithms to apply our framework on real and synthetic datasets.

## 1 Introduction

Predictive models are used in an increasingly high-stakes set of applications, from bail decisions in the criminal justice system [2, 23] to treatment recommendations in personalized medicine [4]. As the stakes have risen, so has the negative impact of incorrect predictions, which could be due to a poorly trained model or to undetected confounding patterns within the data itself [31].

As machine learning models influence a growing fraction of everyday life, individuals often want to understand the reasons for the decisions that affect them. Many governments now recognize a “right to explanation” for significant decisions, for instance as part of the European Union’s General Data Protection Regulation [19]. However, state-of-the-art machine learning methods such as random forests and neural networks are black boxes: their complex structure makes it difficult for humans, including domain experts, to understand their predictive behavior [8, 17].

### 1.1 Interpretable Machine Learning

According to Leo Breiman [9], machine learning has two objectives: prediction, i.e., determining the value of the target variable for new inputs, and information, i.e., understanding the natural relationship between the input features and the target variable. Studies have shown that many decision makers exhibit an inherent distrust of automated predictive models, even if they are proven to be more accurate than human forecasters [12]. One way to overcome “algorithm aversion” is to give decision makers agency to modify the model’s predictions [13]. Another is to provide them with understanding.

Many studies in machine learning seek to train more interpretable models in lieu of complex black boxes. Decision trees [7, 3] are considered interpretable for their discrete structure and graphical visualization, as are close relatives including rule lists [27, 36], decision sets [25], and case-based reasoning [21]. Other approaches include generalized additive models [30], i.e., linear combinations of single-feature models, and score-based methods [35], where integer point values for each feature can be summed up into a final “score”. In the case of linear models, interpretability often comes down to sparsity (small number of nonzero coefficients), a topic of extensive study over the past twenty years [20]. Sparse regression models can be trained using heuristics such as LASSO, stagewise regression or least-angle regression [34, 33, 15], or scalable mixed-integer approaches [5, 6].

Many practitioners are hesitant to give up the high accuracy of black box models in the name of interpretability, and prefer to construct *ex post* explanations for a model’s predictions. Some approaches create a separate explanation for each prediction in the dataset, e.g. by approximating the nonlinear decision boundary of a neural network with a hyperplane [32]. Others define metrics of feature importance to quantify the effect of each feature in the overall model [17, 11].

Finally, some approaches seek to approximate a large, complex model such as a neural network or a random forest with a simpler one – a decision tree [1], two-level rule list [26], or smaller neural network [10]. Such *global* explanations can help human experts detect systemic biases or confounding variables. However, even if these approximations are almost as accurate as the original model, they may have very different behavior on some inputs and can thus provide a misleading assessment of the model’s behavior [18].

The interpretability of linear models and the resulting tradeoff with predictive accuracy are of significant interest to the machine learning community [16]. However, a major challenge in this line of research is that the very concept of interpretability is hard to define and even harder to quantify [29]. Many definitions of interpretability have a “know it when you see it” aspect which impedes quantitative analysis: though some aspects of interpretability are easy to measure, others may be difficult to evaluate without human input [14].

## 1.2 Contributions

We introduce the framework of *interpretable paths*, in which models are decomposed into simple building blocks. An interpretable path is a sequence of models of increasing complexity which can represent a sequential process of “reading” or “explaining” a model. Using examples of several machine learning model classes, we show that the framework of interpretable paths is relevant and intuitively captures properties associated with interpretability.

To formalize which paths are more interpretable, we introduce path interpretability metrics. We define *coherence* conditions that such metrics should satisfy, and derive a parametric family of coherent metrics.

This study of interpretable paths naturally leads to a family of *model* interpretability metrics. The proposed metrics generalize a number of proxies for interpretability from the literature, such as sparsity in linear models and number of splits for decision trees, and also encompass other desirable characteristics.

The model interpretability metrics can be used to select models that are both accurate and interpretable. To this end, we formulate the optimization problem of computing models that are on the Pareto front of interpretability and predictive accuracy (price of interpretability). We give examples in various settings, and discuss computational challenges.

We study an in-depth application to linear models on real and synthetic datasets. We discuss both the modeling aspect (the choice of the interpretability metric), as well as the computational aspect for which we propose exact mixed-integer formulations and scalable local improvement heuristics.

## 2 A Sequential View of Model Construction

### 2.1 Selecting a Model

Most machine learning problems can be viewed through the lens of optimization. Given a set of models  $\mathcal{M}$ , each model  $m \in \mathcal{M}$  is associated with a cost  $c(m) > 0$ , typically derived from data, representing the performance of the model on the task at hand (potentially including a regularization term). Training a machine learning model means choosing the appropriate  $m$  from  $\mathcal{M}$  (for example the one that minimizes  $c(m)$ ). To make this perspective more concrete, we will use the following examples throughout the paper.

**Linear models.** Given the feature matrix  $X \in \mathbb{R}^{n \times d}$  of a dataset of size  $n$  with feature space in  $\mathbb{R}^d$  and the corresponding vector of labels  $y \in \mathbb{R}^n$ , a linear model corresponds to a set of linear coefficients  $\beta \in \mathbb{R}^d$ . In this example,

$\mathcal{M} = \mathbb{R}^d$ , and the cost  $c(\cdot)$  depends on the application: for ordinary least squares (OLS),  $c(\beta) = (1/n)\|X\beta - y\|^2$  (mean squared error).

**Classification trees (CART).** In this case, each model corresponds to a binary decision tree structure [7], so  $\mathcal{M}$  is the set of all possible tree structures of any size. Given a tree  $t \in \mathcal{M}$  and an input  $x \in \mathbb{R}^d$ , let  $t(x)$  designate the tree’s estimate of the corresponding label. Then a typical performance metric  $c(t)$  is the number of misclassified points. If we have a dataset with  $n$  points  $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$  associated with classification labels  $(y_1, \dots, y_n) \in \{0, 1\}^n$  then we have  $c(t) = \sum_{i=1}^n \mathbf{1}(t(x_i) \neq y_i)$ .

**Clustering.** We consider the k-means clustering problem for a dataset  $\mathcal{D}$  of  $n$  points in dimension  $d$ . Our model space  $\mathcal{M}$  is the set of all partitions of the dataset, each partition representing a cluster. Formally  $\mathcal{M} = \cup_{K=1}^n \{(A_1, \dots, A_K) : i \neq j \Rightarrow A_i \cap A_j = \emptyset, \cup_{i=1}^K A_i = \mathcal{D}\}$ . To evaluate a partition, we can use the within-cluster sum of squares  $c(A_1, \dots, A_K) = \sum_{k=1}^K \sum_{x_i \in A_k} \|x_i - \mu_k\|^2$ , where  $\mu_k = \sum_{x_i \in A_k} x_i / |A_k|$  is the centroid of cluster  $A_k$ .

## 2.2 Interpretable Steps

For our guiding examples, typical proxies for interpretability include sparsity in linear models [5], a small number of nodes in a classification tree [3], or a small number of clusters.

As we try to rationalize why these are good proxies for interpretability, one possible approach is to consider how humans read and explain these models. For example, a linear model is typically introduced coefficient by coefficient, a tree is typically read node by node from the root to the leaves, and clusters are typically examined one by one. During this process, we build a model that is more and more complex. In other words, the human process of understanding a model can be viewed as a decomposition into simple building blocks.

We introduce the notion of an *interpretable step* to formalize this sequential process. For every model  $m \in \mathcal{M}$ , we define a step neighborhood function  $\mathcal{S}$  that associates each model  $m$  to the set of models  $\mathcal{S}(m) \subseteq \mathcal{M}$  such that  $m'$  is one interpretable step away from  $m$  if and only if  $m' \in \mathcal{S}(m)$ . Interpretable steps represent simple model updates that can be chained to build increasingly complex models.

For linear models, one possible interpretable step is modifying a single coefficient, i.e.  $\beta'$  belongs to  $\mathcal{S}(\beta)$  if  $\|\beta - \beta'\|_0 \leq 1$  ( $\beta$  and  $\beta'$  differ in at most one coefficient). For CART, an interpretable step could be adding a split to an existing tree, i.e.,  $t' \in \mathcal{S}(t)$  if  $t'$  can be obtained by splitting a leaf node of  $t$  into two leaves. For clustering, we could use the structure of hierarchical clustering and choose a step that increases the number of clusters by one by splitting an existing cluster into two. These examples are illustrated in Figure 1.

Choosing the step neighborhood function  $\mathcal{S}$  is a modeling choice and for the examples considered, there may be many other ways to define it. To simplify the analysis, we only impose that  $\mathcal{S}(m) \neq \emptyset$  for all  $m$  (there must always be a feasible next step from any model), which can trivially be satisfied by ensuring  $m \in \mathcal{S}(m)$  (an interpretable step can involve no changes to the model).

Given the choice of an interpretable step  $\mathcal{S}$ , we can define an *interpretable path* of length  $K$  as a sequence of  $K$  models  $\mathbf{m} = (m_1, \dots, m_K)$  such that  $m_k \in \mathcal{S}(m_{k-1})$  for all  $1 \leq k \leq K$ , i.e., a sequence of interpretable steps starting from a base model  $m_0$ . The choice of  $m_0$ , the “simplest” model, is usually obvious: in our examples,  $m_0$  could be a linear model with  $\beta = \mathbf{0}$ , an empty classification tree, or a single cluster containing all data points. Given the model space  $\mathcal{M}$ , we call  $\mathcal{P}_K$  the set of all interpretable paths of length  $K$  and  $\mathcal{P} = \cup_{K=0}^{\infty} \mathcal{P}_K$  the set of all interpretable paths of any length.

Let us consider an example with classification trees to build intuition about interpretable paths. The *iris* dataset is a small dataset often used to illustrate classification problems. It records the petal length and width and sepal length and width of various iris flowers, along with their species (*setosa*, *versicolor* and *virginica*). For simplicity, we only consider two of the four features (petal length and width) and subsample 50 total points from two of the three classes (*versicolor* and *virginica*).

We define an interpretable step as splitting one leaf node into two. Given the *iris* dataset, we consider two classification trees  $t_{\text{good}}, t_{\text{bad}} \in \mathcal{M}$ . Both trees have a depth of 2, exactly 3 splits, and a misclassification cost of 2. However, when we consider interpretable paths leading to these two trees, we notice some differences. An interpretable path  $\mathbf{t}_{\text{bad}} \in \mathcal{P}$  leading to  $t_{\text{bad}} \in \mathcal{M}$  is shown in Figure 2, and an interpretable path  $\mathbf{t}_{\text{good}}$  leading to  $t_{\text{good}}$  is detailed in Figure 3.

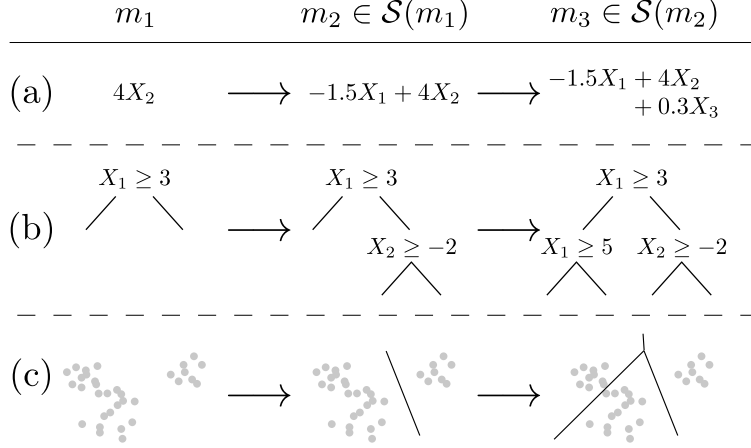


Figure 1: Illustration of the interpretable path framework with the three examples introduced in Section 2.1: (a) is our linear model setting; (b) corresponds to the classification trees (CART); (c) to the clustering setting (in 2 dimensions). For each space of model, we illustrate an example of interpretable path following the choice of steps introduced in Section 2.2. Each path has 3 steps:  $m_1$ ,  $m_2$  and  $m_3$ .

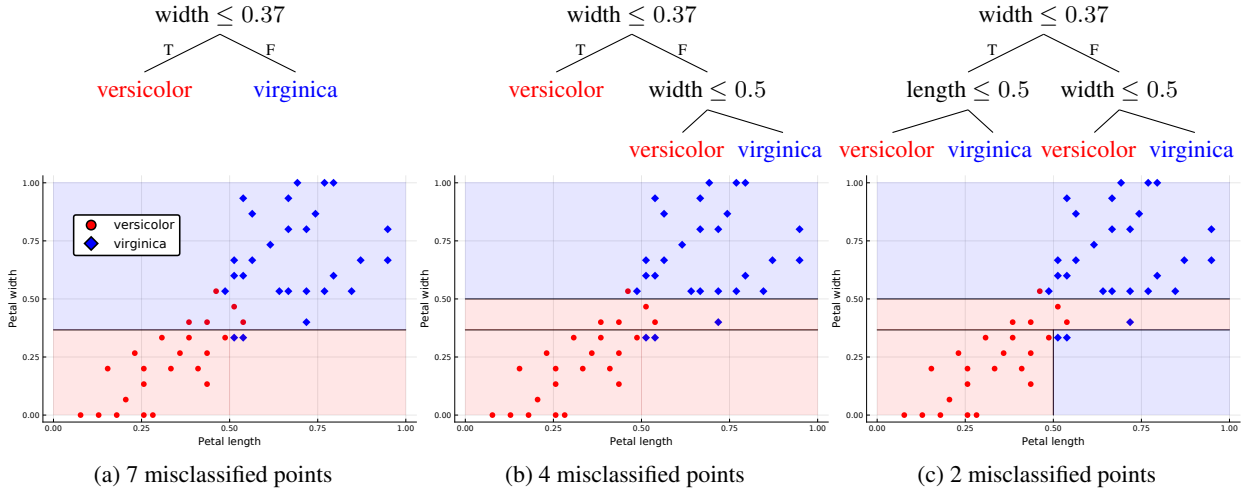


Figure 2: Visualization of an interpretable path leading to  $t_{\text{bad}}$  (shown on the right).

For  $t_{\text{bad}}$ , the first split results in an intermediate tree with a high classification error, which could be less intuitive for flower connoisseurs. In contrast, the first split of  $t_{\text{good}}$  gives a much more accurate intermediate tree. We will introduce a way to formally identify which of the two paths is more interpretable.

### 3 The Tradeoffs of Interpretability

In this section, we consider the choice of an interpretability loss  $\mathcal{L}(m)$  for all interpretable paths  $m \in \mathcal{P}$  such that a path  $m$  is considered more interpretable than a path  $m'$  if and only if  $\mathcal{L}(m) < \mathcal{L}(m')$ . We first motivate this formalism by showing it can lead to a notion of interpretability loss for models as well. We then use a simple example to build intuition about the choice of  $\mathcal{L}$ .

#### 3.1 From paths to models

Defining a loss function for the interpretability of a path can naturally lead to an interpretability loss on the space of models, with the simple idea that more interpretable paths should lead to more interpretable models. Given a path

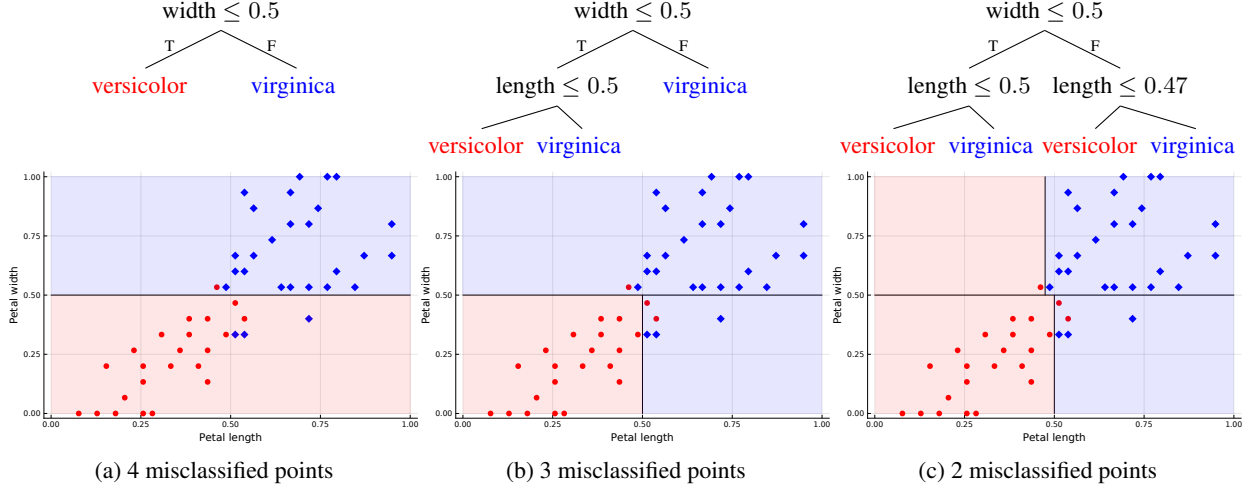


Figure 3: Visualization of an interpretable path leading to  $t_{\text{good}}$  (shown on the right).

interpretability loss  $\mathcal{L}(\cdot)$ , we can define a corresponding model interpretability loss as

$$\mathcal{L}(m) = \begin{cases} \infty, & \text{if } \mathcal{P}(m) = \emptyset, \\ \min_{\mathbf{m} \in \mathcal{P}(m)} \mathcal{L}(\mathbf{m}), & \text{otherwise,} \end{cases} \quad (1)$$

where  $\mathcal{P}_K(m) = \{\mathbf{m} \in \mathcal{P}_K, m_K = m\}$  designates the set of interpretable paths of length  $K$  leading to  $m$ , and  $\mathcal{P}(m) = \cup_{K=0}^{\infty} \mathcal{P}_K(m)$  designates the set of finite interpretable paths leading to  $m$ . In other words, the interpretability loss of a model  $m$  is the interpretability loss of the most interpretable path leading to  $m$ .

As an example, consider the following path interpretability loss, which we call *path complexity* and define as  $\mathcal{L}_{\text{complexity}}(\mathbf{m}) = |\mathbf{m}|$  (number of steps in the path). Under this metric, paths are considered less interpretable if they are longer. From (1) we can then define the interpretability loss of a given model  $m$  as

$$\mathcal{L}_{\text{complexity}}(m) = \min_{\mathbf{m} \in \mathcal{P}(m)} |\mathbf{m}|,$$

which corresponds to the minimal number of interpretable steps required to reach  $m$ .

In the context of the examples from Section 2, the function  $\mathcal{L}_{\text{complexity}}$  recovers typical interpretability proxies. For a linear model  $\beta$ ,  $\mathcal{L}_{\text{complexity}}(\beta) = \|\beta\|_0$  is the sparsity of the model (number of non-zero coefficients). For a classification tree  $t$ ,  $\mathcal{L}_{\text{complexity}}(t)$  is the number of splits. In a clustering context,  $\mathcal{L}_{\text{complexity}}(A_1, \dots, A_K) = K$  is just the number of clusters. We refer to this candidate loss function as the *model complexity*.

A fundamental problem of interpretable machine learning is finding the highest-performing model at a given level of interpretability [9]. Defining an interpretability loss function  $\mathcal{L}(\cdot)$  on the space of models  $\mathcal{M}$  is important because it allows us to formulate this problem generally as follows:

$$\min_{m \in \mathcal{M}} c(m) \quad \text{s.t.} \quad \mathcal{L}(m) \leq \ell, \quad (2)$$

where  $\ell$  is the desired level of interpretability. Problem (2) produces models on the Pareto front of accuracy and interpretability. If we compute this Pareto front by solving problem (2) for any  $\ell$ , then we can mathematically characterize the *price* of our definition of interpretability on our dataset given a class of models, making the choice of a final model easier.

In the case of model complexity  $\mathcal{L}_{\text{complexity}}$ , for  $\ell = K$  problem (2) can be written as

$$\min_{\mathbf{m} \in \mathcal{P}_K} c(\mathbf{m}_K). \quad (3)$$

Problem (3) generalizes existing problems in interpretable machine learning: best subset selection ( $L_0$ -constrained sparse regression) for linear models [5], finding the best classification tree of a given size [3], or the K-means problem of finding the  $K$  best possible clusters.

Thus, the framework of interpretable paths naturally gives rise to a general definition of model complexity via the loss function  $\mathcal{L}_{\text{complexity}}$ , and our model generalizes many existing approaches. By some counts, however, model

complexity remains an incomplete interpretability loss. For instance, it does not differentiate between the trees  $t_{\text{good}}$  and  $t_{\text{bad}}$ : both models have a complexity of 3 because they can be reached in three steps. More generally,  $\mathcal{L}_{\text{complexity}}$  does not differentiate between paths of the same length, or between models that can be reached by paths of the same length.

### 3.2 Incrementality

In the decision tree example from Figures 2 and 3, we observed that the intermediate trees leading to  $t_{\text{good}}$  were more accurate than the intermediate trees leading to  $t_{\text{bad}}$ . Evaluating the costs  $c(m_k)$  of intermediate models along a path  $\mathbf{m}$  may provide clues as to the interpretability of the final model.

Consider the following toy example, where the goal is to estimate a child’s age  $y_{\text{Age}}$  given height  $X_{\text{Height}}$  and weight  $X_{\text{Weight}}$ . The normalized features  $X_{\text{Height}}$  and  $X_{\text{Weight}}$  have correlation  $\rho = 0.9$  and are both positively correlated with the objective. Solving the OLS problem yields  $\beta^* = (2.12, -0.94)$ , i.e.,

$$y_{\text{Age}} = 2.12 \cdot X_{\text{Height}} - 0.94 \cdot X_{\text{Weight}} + \varepsilon, \tag{4}$$

with  $\varepsilon = X\beta^* - y$  the error term. The mean squared error (MSE) of the model  $\beta^*$  is  $c(\beta^*) = \frac{1}{n} \sum_i \varepsilon_i^2 = 0.25$ .

As in Section 2, we define an interpretable step to be modifying a single coefficient in the linear model, keeping all other coefficients constant. In this case, consider the three interpretable paths in Table 1. When using the complexity loss  $\mathcal{L}_{\text{complexity}}$ , the first two paths in the table are considered equally interpretable because they have the same length. But are they? Both verify  $c(m_2) = c(m'_2) = 0.25$ , but  $c(m_1) = 1.13 < c(m'_1) = 4.74$ . Indeed,  $m'_1$  is a particularly inaccurate model, as weight is positively correlated with age. And furthermore, if having an accurate first step matters to the user, then path  $\hat{\mathbf{m}}$  may be preferred even though it is longer.

| $\mathbf{m}$          | $c(m_i)$ | $\mathbf{m}'$          | $c(m'_i)$ | $\hat{\mathbf{m}}$          | $c(\hat{m}_i)$ |
|-----------------------|----------|------------------------|-----------|-----------------------------|----------------|
| $m_0 = (0, 0)$        | 2.04     | $m'_0 = (0, 0)$        | 2.04      | $\hat{m}_0 = (0, 0)$        | 2.04           |
| $m_1 = (2.12, 0)$     | 1.13     | $m'_1 = (0, -0.94)$    | 4.74      | $\hat{m}_1 = (1.70, 0)$     | 0.60           |
| $m_2 = (2.12, -0.94)$ | 0.25     | $m'_2 = (2.12, -0.94)$ | 0.25      | $\hat{m}_2 = (1.70, -0.94)$ | 0.43           |
|                       |          |                        |           | $\hat{m}_3 = (2.12, -0.94)$ | 0.25           |

(a) Two steps, starting with height

(b) Two steps, starting with weight

(c) Three steps

Table 1: Three decompositions of  $m^*$  into a sequence of interpretable steps.

As discussed in Section 2, an interpretable path  $\mathbf{m}$  leading to model  $m$  can be viewed as a decomposition of  $m$  into a sequence of easily understandable steps. The costs of intermediate models should play a role in quantifying the interpretability loss of a path; higher costs should be penalized, as we want to avoid nonsensical intermediate models such as  $m'_1$ .

One way to ensure that every step of an interpretable path adds value is a greedy approach, where the next model at each step is chosen by minimizing the cost  $c(\cdot)$ :

$$m_{k+1}^{\text{greedy}} \in \arg \min \left\{ c(m), m \in \mathcal{S}(m_k^{\text{greedy}}) \right\} \quad \forall k \geq 1. \tag{5}$$

In our toy example, restricting ourselves to paths of length 2, this means selecting the best possible  $m_1^{\text{greedy}}$ , and then the best possible  $m_2^{\text{greedy}}$  given  $m_1^{\text{greedy}}$ , as in stagewise regression [33]. This will not yield the best possible model achievable in two steps as in (3), but the first step is guaranteed to be the best one possible. Notice that  $c(m_1^{\text{greedy}}) = 0.42 < 1.13 = c(m_1)$ , but  $c(m_2^{\text{greedy}}) = 0.39 > 0.25 = c(m_2)$ . The improvement of the first model comes at the expense of the second step.

Deciding which of the two paths  $\mathbf{m}$  and  $\mathbf{m}^{\text{greedy}}$  is more interpretable is a hard question. It highlights the tradeoff between the desirable incrementality of the greedy approach and the cost of the final model. For paths of length 2, there is a continuum of models between  $\mathbf{m}$  and  $\mathbf{m}^{\text{greedy}}$ , corresponding to the Pareto front between  $c(m_1)$  and  $c(m_2)$ , shown in Figure 4.

## 4 Coherent Interpretability Losses

In the previous section, we developed intuition regarding the interpretability of different paths. We now formalize this intuition in order to define a suitable interpretability loss.

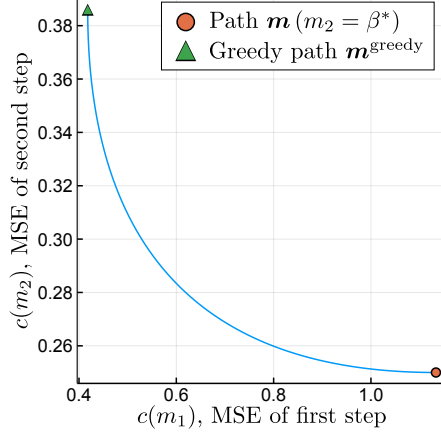


Figure 4: Tradeoff between the cost of the first and second models of the interpretable path.

#### 4.1 Coherent Path Interpretability Losses

According to the loss  $\mathcal{L}_{\text{complexity}}$  defined in Section 3.1, which generalizes many notions of interpretability from the literature, a path is more interpretable if it is shorter. In Section 3.2, we saw that the cost of individual models along the path matters as well.

Sometimes, comparing the costs of intermediate models between two paths is easy because the cost of each step along one path is at least as good as the cost of the corresponding step in the other path. In Table 1, it is reasonable to consider  $\mathbf{m}$  more interpretable than  $\mathbf{m}'$  because  $c(m_1) < c(m'_1)$  and  $c(m_2) = c(m'_2)$ . In contrast, comparing the interpretability of  $\mathbf{m}$  and  $\mathbf{m}^{\text{greedy}}$  is more difficult and user-specific, because  $c(m_1) > c(m_1^{\text{greedy}})$ , but  $c(m_2) < c(m_2^{\text{greedy}})$ .

We now formalize this intuition into desirable properties of interpretability loss functions. We first introduce the notion of a cost sequence, which provides a concise way to refer to the costs of all the steps in an interpretable path. We then propose axioms for *coherent* interpretability losses.

**Definition 1** (Cost sequence). Given an interpretable path of length  $K$ , denoted as  $\mathbf{m} \in \mathcal{P}_K$ , the cost sequence  $\mathbf{c}(\mathbf{m}) \in \mathbb{R}^{\mathbb{N}}$  is the infinite sequence  $(c_1, c_2, \dots)$  such that:

$$c_k = \begin{cases} c(m_k), & \text{if } k \leq K, \\ 0, & \text{otherwise.} \end{cases}$$

**Definition 2** (Coherent Interpretability Loss). A path interpretability loss  $\mathcal{L}$  is *coherent* if the following conditions hold for any two interpretable paths  $\mathbf{m}, \mathbf{m}' \in \mathcal{P}$  with respective cost sequences  $\mathbf{c}$  and  $\mathbf{c}'$ .

- (a) If  $\mathbf{c} = \mathbf{c}'$ , then  $\mathcal{L}(\mathbf{m}) = \mathcal{L}(\mathbf{m}')$ .
- (b) (Weak Pareto dominance) If  $c_k \leq c'_k \forall k$  (which we write as  $\mathbf{c} \leq \mathbf{c}'$ ), then  $\mathcal{L}(\mathbf{m}) \leq \mathcal{L}(\mathbf{m}')$ .

Condition (a) means that the interpretability of a path depends only on the sequence of costs along that path. Condition (b) formalizes the intuition described before, that paths with fewer steps or better steps are more interpretable. For instance, if we improve the cost of one step of a path while leaving all other steps unchanged, we can only make the path more interpretable. Under any coherent interpretability loss  $\mathcal{L}$  in Table 1,  $\mathbf{m}$  is more interpretable than  $\mathbf{m}'$ , but  $\mathbf{m}$  may be more or less interpretable than  $\mathbf{m}^{\text{greedy}}$  depending on the specific choice of coherent interpretability loss.

In addition, consider a path  $\mathbf{m} \in \mathcal{P}_K$  and remove its last step to obtain a new path  $\mathbf{m}' \in \mathcal{P}_{K-1}$ . This is equivalent to setting the  $K$ -th element of the cost sequence  $\mathbf{c}(\mathbf{m})$  to zero. Since  $c(\cdot) > 0$ , we have that  $\mathbf{c}(\mathbf{m}') \leq \mathbf{c}(\mathbf{m})$ , which implies  $\mathcal{L}(\mathbf{m}') \leq \mathcal{L}(\mathbf{m})$ . In other words, under a coherent interpretability loss, removing a step from an interpretable path can only make the path more interpretable.

*Remark.* The path complexity  $\mathcal{L}_{\text{complexity}}(\mathbf{m}) = |\mathbf{m}|$  is a coherent path interpretability loss.

*Proof.* If  $\mathbf{m}$  and  $\mathbf{m}'$  verify  $\mathbf{c}(\mathbf{m}) = \mathbf{c}(\mathbf{m}')$ , then trivially the two cost sequences become zero after the same number of steps, so  $\mathcal{L}_{\text{complexity}}(\mathbf{m}) = \mathcal{L}_{\text{complexity}}(\mathbf{m}')$ . If  $\mathbf{c}(\mathbf{m}) \leq \mathbf{c}(\mathbf{m}')$  and  $\mathbf{c}(\mathbf{m}')$  becomes zero after exactly  $K$  steps, then  $\mathbf{c}(\mathbf{m})$  must become zero after at most  $K$  steps, so  $\mathcal{L}_{\text{complexity}}(\mathbf{m}) \leq \mathcal{L}_{\text{complexity}}(\mathbf{m}')$ .  $\square$

## 4.2 A Coherent Model Interpretability Loss

Axiom (b) of Definition 2 states that a path that dominates another path in terms of the costs of each step must be at least as interpretable. This notion of weak Pareto dominance suggests a natural path interpretability loss:

$$\mathcal{L}_\alpha(\mathbf{m}) = \alpha \cdot \mathbf{c}(\mathbf{m}) = \sum_{k=1}^{|\mathbf{m}|} \alpha_k c(m_k).$$

In other words, the interpretability loss  $\mathcal{L}_\gamma$  of a path  $\mathbf{m}$  is the weighted sum of the costs of all steps in the path. This loss function is trivially coherent and extremely general. It is parametrized by the infinite sequence of weights  $\alpha = (\alpha_1, \alpha_2, \dots)$ , which specifies the relative importance of the accuracy of each step in the model for the particular application at hand.

Defining a family of interpretability losses with infinitely many parameters allows for significant modeling flexibility, but it is also cumbersome and overly general. We therefore propose to select  $\alpha_k = \gamma^k$  for all  $k$ , replacing the infinite sequence of parameters  $(\alpha_1, \alpha_2, \dots)$  with a single parameter  $\gamma > 0$ . In this case, following (1), we propose the following coherent interpretability loss function on the space of models.

**Definition 3** (Model interpretability). Given a model  $m \in \mathcal{M}$ , its interpretability loss  $\mathcal{L}_\gamma(m)$  is given by

$$\mathcal{L}_\gamma(m) = \begin{cases} \infty, & \text{if } \mathcal{P}(m) = \emptyset, \\ \min_{\mathbf{m} \in \mathcal{P}(m)} \mathcal{L}_\gamma(\mathbf{m}) = \sum_{k=1}^{|\mathbf{m}|} \gamma^k c(m_k), & \text{otherwise.} \end{cases} \quad (6)$$

By definition,  $\mathcal{L}_\gamma$  is a coherent interpretability loss, which favors more incremental models or models with a low complexity. The parameter  $\gamma$  captures the tradeoff between these two aspects of interpretability. Theorem 1 shows that with a particular choice of  $\gamma$  one can recover the notion of model complexity introduced in Section 3.1, or models that can be built in a greedy way.

**Theorem 1** (Consistency of interpretability measure). *Assume that the cost  $c(\cdot)$  is bounded, we consider  $\mathcal{L}_\gamma(m)$  in the two limit cases  $\gamma \rightarrow +\infty$  and  $\gamma \rightarrow 0$ :*

- (a) *Let  $m^+, m^- \in \mathcal{M}$  with  $\mathcal{L}_{\text{complexity}}(m^+) < \mathcal{L}_{\text{complexity}}(m^-)$  (i.e.,  $m^+$  requires less interpretable steps than  $m^-$ ), or  $\mathcal{L}_{\text{complexity}}(m^+) = \mathcal{L}_{\text{complexity}}(m^-)$  and  $c(m^+) < c(m^-)$ .*

$$\lim_{\gamma \rightarrow \infty} \mathcal{L}_\gamma(m^-) - \mathcal{L}_\gamma(m^+) = +\infty. \quad (7)$$

- (b) *Given  $m^+, m^- \in \mathcal{P}$ , if  $c(m^+) \preceq c(m^-)$ , where  $\preceq$  represents the lexicographic order on  $\mathbb{R}^{\mathbb{N}}$ , then*

$$\lim_{\gamma \rightarrow 0} \mathcal{L}_\gamma(m^-) - \mathcal{L}_\gamma(m^+) \geq 0. \quad (8)$$

*Consequently, given models  $m^+, m^- \in \mathcal{M}$ , if there is  $m^+ \in \mathcal{P}(m^+)$  such that  $c(m^+) \preceq c(m^-)$  for all  $m^- \in \mathcal{P}(m^-)$ , then*

$$\lim_{\gamma \rightarrow 0} \mathcal{L}_\gamma(m^-) - \mathcal{L}_\gamma(m^+) \geq 0. \quad (9)$$

Intuitively, in the limit  $\gamma \rightarrow +\infty$ , (a) states that the most interpretable models are the ones with minimal complexity, or minimal costs if their complexity is the same. (b) states that in the limit  $\gamma \rightarrow 0$  the most interpretable models can be constructed with greedy steps. Definition 3 therefore generalizes existing approaches and provides a good framework to model the tradeoffs of interpretability.

## 5 Interpretability Losses in Practice

Defining an interpretability loss brings a new perspective to the literature on interpretability in machine learning. In this section, we discuss the applications of this framework. For the sake of generality, in the early part of this section we work with the more general interpretability loss  $\mathcal{L}_\alpha(\cdot)$ .



## 5.1 The Price of Interpretability

Given the metric of interpretability defined above, we can quantitatively discuss the price of interpretability, i.e., the tradeoff between a model's interpretability loss  $\mathcal{L}_\alpha(m)$  and its cost  $c(m)$ . To evaluate this tradeoff, we want to compute models that are Pareto optimal with respect to  $c(\cdot)$  and  $\mathcal{L}_\alpha(\cdot)$ , as in (2).

Computing these Pareto-optimal solutions can be challenging, as our definition of model interpretability requires optimizing over paths of any length. Fortunately, the only optimization problem we need to be able to solve is to find the most interpretable path of a fixed length  $K$ , i.e.,

$$\min_{\mathbf{m} \in \mathcal{P}_K} \mathcal{L}_\alpha(\mathbf{m}) = \sum_{k=1}^K \alpha_k c(m_k) \quad (10)$$

Indeed, the following proposition shows that we can compute Pareto-optimal solutions by solving a sequence of optimization problems (10) for various  $K$  and  $\alpha$ .

**Proposition 1** (Price of interpretability). *Pareto-optimal models that minimize the interpretability loss  $\mathcal{L}_\alpha$  and the cost  $c(\cdot)$  can be computed by solving the following optimization problem:*

$$\min_{K \geq 0} \left( \min_{\mathbf{m} \in \mathcal{P}_K} c(m_K) + \lambda \sum_{k=1}^K \alpha_k c(m_k) \right), \quad (11)$$

where  $\lambda \in \mathcal{R}$  is a tradeoff parameter between cost and interpretability.

The (simple) proof of the proposition is provided in the appendix. Notice that the inner minimization problem in (11) is simply problem (10) with the modified coefficients  $(\lambda\alpha_1, \dots, \lambda\alpha_{K-1}, (1+\lambda)\alpha_K)$ .

By defining the general framework of coordinate paths and a natural family of coherent interpretability loss functions, we can understand exactly how much we gain or lose in terms of accuracy when we choose a more or less interpretable model. Our framework thus provides a principled way to answer a central question of the growing literature on interpretability in machine learning.

Readers will notice that the weighted sum of the objectives optimized in Proposition 1 does not necessarily recover the entire Pareto front, and in particular cannot recover any non-convex parts [22].

Using Proposition 1, we can compute the price of interpretability for a range of models and interpretability losses. As an example, Figure 5 shows all Pareto-optimal models with respect to performance cost and interpretability for our toy problem from Section 3.2, with the interpretability loss  $\mathcal{L}_\gamma$  chosen such that  $\gamma = 1$ . Figure 6 shows the Pareto front in the same setting for other values of  $\gamma$ . We notice that as in Theorem 1, when  $\gamma$  grows large our notion of interpretability reduces to sparsity (discrete Pareto curve), whereas when  $\gamma$  grows small our notion of interpretability favors a larger number of incremental steps.

## 5.2 Computational Considerations

To solve (11) we consider a sequence of problems of type (10). However, this sequence is possibly infinite, which poses a computational problem. Proposition 2 provides a bound for the number of problems of type (10) we need to consider in the general case.

**Proposition 2.** *Assume there exist  $c_{\min}$  and  $c_{\max}$  such that  $0 < c_{\min} \leq c(m) \leq c_{\max}$  for all  $m \in \mathcal{M}$  (positive and bounded cost function), and consider the interpretability loss  $\mathcal{L}_\gamma$ . If  $\gamma \geq 1$ , then*

$$K_{opt} := \arg \min_{K \geq 0} \left( \min_{\mathbf{m} \in \mathcal{P}_K} c(m_K) + \lambda \sum_{k=1}^K \gamma^k c(m_k) \right) \leq K_{\max}, \quad (12)$$

where

$$K_{\max} = \begin{cases} \frac{c_{\max}}{\lambda c_{\min}} & \text{if } \gamma = 1, \\ \frac{\log\left(1 + \frac{(\gamma-1)c_{\max}}{\lambda \gamma c_{\min}}\right)}{\log \gamma} & \text{if } \gamma > 1. \end{cases} \quad (13)$$

In other words, under the interpretability loss  $\mathcal{L}_\gamma$  with  $\gamma \geq 1$ , we can find the optimal solution of (11) by solving at most  $K_{\max}$  problems of type (10). The proof of Proposition 2 is provided in the appendix.

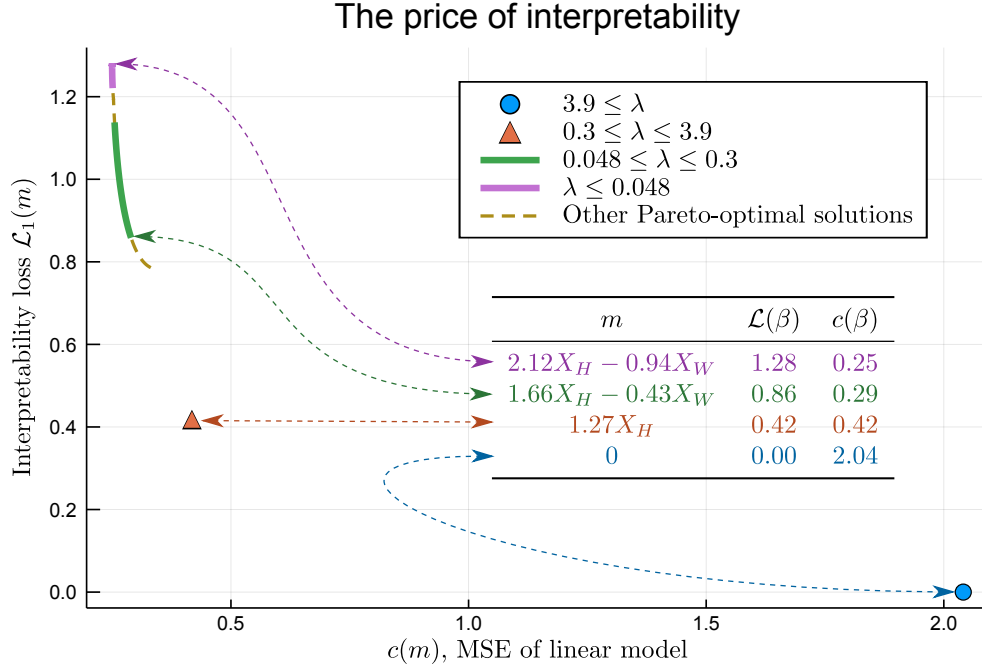
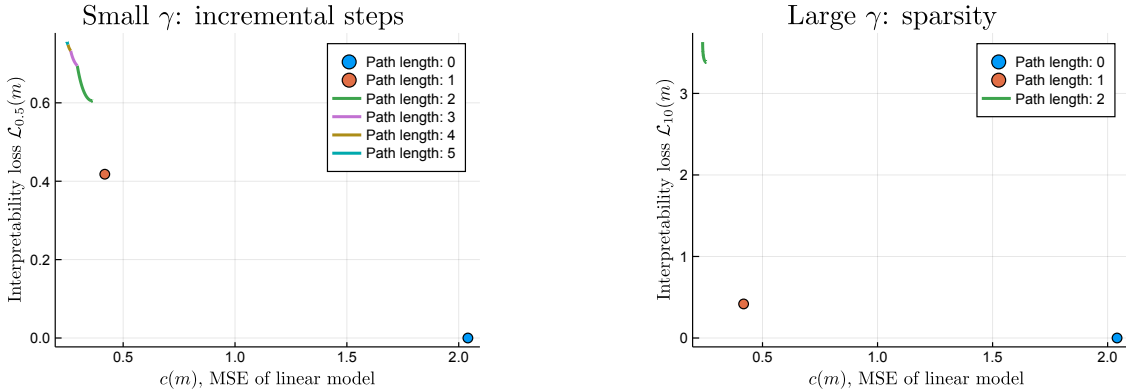


Figure 5: Pareto front between interpretability loss  $\mathcal{L}(m) = \mathcal{L}_\gamma(m)$  (with  $\gamma = 1$ ) and cost  $c(m)$  on the toy OLS problem (4), computed by varying  $\lambda$  in (11). The dashed line represents Pareto-optimal solutions that cannot be computed by this weighted-sum method. Note that the front is discontinuous, and that there is an infinite number of Pareto-optimal models with two steps, but only one respectively with one and zero steps. The inset table describes several interesting Pareto-optimal models.



(a) Here we choose  $\gamma = 0.5$ , therefore the first steps are the most important, and we favor incremental/greedy models, with potentially many steps (see second part of Theorem 1). Each color corresponds roughly to the addition of a greedy step (this becomes exact when  $\gamma \rightarrow 0$ ).

(b) Here we choose  $\gamma = 10$ , therefore the cost of the intermediate steps is much less important than the final models, and we favor sparse models (see first part of Theorem 1). Note that the Pareto front has three almost discrete parts, corresponding to the three possible levels of sparsity in this example.

Figure 6: Pareto fronts between model interpretability and cost in the same setting as Figure 5, except that we change the definition of interpretability by changing the value  $\gamma$ .

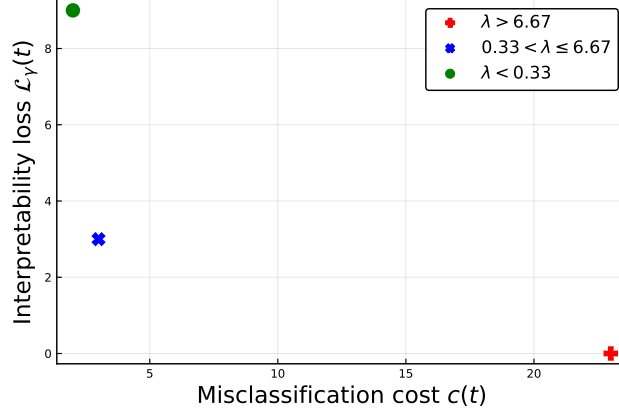


Figure 7: Price of interpretability for decision trees of depth at most 2 on the simplified `iris` dataset.

A corollary of Proposition 2 is that we can write an optimization formulation of problem (11) with a finite number of decision variables. For instance, we can formulate the inner minimization problem with finitely many decision variables for each  $K$  and then solve finitely many such problems. The tractability of this optimization problem is application-dependent.

For example, by adapting the mixed-integer optimization formulation from Bertsimas and Dunn [3], we can compute the price of interpretability for decision trees of bounded depth by writing the following mixed-integer formulation of the inner minimization problem in (11):

$$\min \sum_{k=1}^K \gamma^k f(d_t^k, a_t^k, b_t^k) \quad (14a)$$

$$\text{s.t. } (d_t^k, a_t^k, b_t^k) \in \mathcal{T} \quad \forall k \in [K] \quad (14b)$$

$$\sum_{t \in \mathcal{T}_L} d_t^k = k \quad \forall k \in [K] \quad (14c)$$

$$a_t^k \leq a_t^{k+1} \quad \forall t \in \mathcal{T}_B, k \in [K-1] \quad (14d)$$

$$d_t^k \leq d_t^{k+1} \quad \forall t \in \mathcal{T}_B, k \in [K-1] \quad (14e)$$

$$b_t^k - (1 - d_t^k) \leq b_t^{k+1} \leq b_t^k + (1 - d_t^k) \quad \forall t \in \mathcal{T}_B, k \in [K-1], \quad (14f)$$

where the variables  $d_t^k$ ,  $a_t^k$  and  $b_t^k$  define  $K$  trees of depth at most  $D$ , and constraints (14c)-(14f) impose an interpretable path structure on the  $K$  trees. The set  $\mathcal{T}_B$  indicates the set of branching nodes of the trees, the variable  $d_t^k$  indicates whether branching node  $t$  in tree  $k$  is active,  $a_t^k$  selects the variable along which to perform the split at branching node  $t$  in tree  $k$ , and  $b_t^k$  is the split value at branching node  $t$  in tree  $k$ . The function  $f$  is the objective value of the tree defined by these split variables, and the set  $\mathcal{T}$  designates all the constraints to impose the tree structure for each  $k$  (constraint (14b) is equivalent to (24) from [3]). Constraint (14c) imposes that tree  $k$  must have exactly  $k$  active splits, Constraint (14e) forces tree  $k+1$  to keep all the branching nodes of tree  $k$ , and constraints (14d) and (14f) force the splits at these common branching nodes to be the same.

This formulation allows us to compute the price of interpretability on the simplified `iris` dataset from Section 2. The resulting Pareto curve is shown in Figure 7. It turns out the most interpretable tree with a misclassification error of 2 is  $t_{\text{good}}$ , with  $\mathcal{L}_\gamma(t_{\text{good}}) = 9$  (for  $\gamma = 1$ ).

In general, mixed-integer optimization formulations such as (14) may not scale. However, in many cases a provably optimal solution is not necessary and scalable heuristics such as local improvement may be employed. We provide such an example in Section 6.

### 5.3 Interpretable Paths and Human-in-the-Loop Analytics

Motivated by the idea that humans read and explain models sequentially, we have used the framework of interpretable paths to evaluate the interpretability of individual models. Viewing an interpretable path as a nested sequence of models of increasing complexity can also be useful in the context of human-in-the-loop analytics.

Consider the problem of customer segmentation via clustering. Choosing the number of customer types ( $k$ ) is not always obvious in practice and has to be selected by a decision-maker. Solving the clustering problem with  $k$  clusters and with  $k + 1$  clusters may lead to very different clusters. Alternatively, using interpretable steps, we can force a hierarchical structure on the clusters, i.e., the solution with  $k + 1$  clusters results from the splitting of one of the clusters of the solution with  $k$  clusters, for all  $k$ . The change between  $k$  clusters and  $k + 1$  clusters becomes simpler and may facilitate the choice of  $k$ .

If we assume each  $k$  can be chosen with equal probability for  $k \leq 10$ , the problem of finding the sequence that minimizes the expected cost is:

$$\min_{\mathbf{m} \in \mathcal{P}_{10}} \frac{1}{10} \sum_{k=1}^{10} c(m_k), \quad (15)$$

which is exactly the decision problem (10) with the weights  $\alpha_k = 0.1$  for  $k \leq 10$ , and  $\alpha_k = 0$  otherwise. This problem is related to studies in incremental approximation algorithms [28] and prioritization [24], which are typically motivated by a notion of interpretability which simplifies implementation for practitioners.

More generally, we can use interpretable paths to facilitate human-in-the-loop model selection. Given a discrete distribution on the choice of  $K$ :  $p_k = \mathbb{P}(\{k \text{ will be chosen by the decision maker}\})$ , we can choose  $\alpha_k = p_k$  and solve (10) to find paths  $\mathbf{m}$  that minimize the expected cost  $\mathbb{E}_k[c(m_k)]$ .

## 6 Application: Linear Regression

So far, we have presented a mathematical framework to formalize the discussion of interpretability. We now study in detail how it can be used in practice, focusing on the single application of linear regression.

### 6.1 Modeling interpretability

In the example of linear regression, we defined the following interpretable steps:

$$\mathcal{S}(\beta) = \{\beta' \in \mathbb{R}^d : \|\beta - \beta'\|_0 \leq 1\}. \quad (16)$$

These steps are a modeling choice. They lead to decompositions of linear models where the coefficients are introduced or modified one at a time, and we have seen they are intimately linked to sparsity. We wish to obtain models that can easily be introduced coefficient by coefficient, allowing ourselves to modify coefficients that have already been set.

Choosing a different step function  $\mathcal{S}(\cdot)$  can lead to other notions of interpretability. For instance, each step could add a feature, allowing to modify all the weights (not only one coordinate):  $\mathcal{S}_{\text{features}}(\beta) = \{\theta \text{ s.t. } \|\theta\|_0 \leq \|\beta\|_0 + 1\}$ . This boils down to ordering the features of a linear model, finding the most interpretable order. We could also choose  $\mathcal{S}_{\text{SLIM}}(\beta) = \{\theta : \|\beta - \theta\|_0 \leq 1, \|\beta - \theta\|_1 \in \mathbb{Z}\}$ , which imposes integer coordinate updates at each step. This is related to the notion of interpretability introduced by score-based methods [35]. Another way to think about score-based methods is to choose  $\mathcal{S}'_{\text{SLIM}}(\beta) = \{\theta : \|\beta - \theta\|_0 \leq 1, \|\beta - \theta\|_1 \in \{0, 1\}\}$ , which imposes that each step adds one point to the scoring system.

We select the interpretability loss  $\mathcal{L}_\gamma$  with  $\gamma = 1$  (meaning the costs of all steps matter equally). Given the step function  $\mathcal{S}$  defined in (16), the convex quadratic cost function  $c(\cdot)$ , and the initial regression coefficients  $\beta_0$ , as in Section 5, our goal is to find the optimal interpretable path of length  $K$  (10).

### 6.2 Algorithms

**Optimal.** Problem (10) can be written as a convex integer optimization problem using special ordered sets of type 1 (SOS-1 constraints).

$$\min_{\beta_k} \sum_{k=1}^K c(\beta_k) \quad (17a)$$

$$\text{s.t. SOS-1}(\beta_{k+1} - \beta_k) \quad 0 \leq k < K. \quad (17b)$$

For reasonable problem sizes ( $d \leq 10$ ,  $K \leq 10$  and any choice of  $n$ ), this problem can be solved exactly using a standard solver such as Gurobi or CPLEX.

**Local improvement.** In higher-dimensional settings, or when  $K$  is too large, the formulation above may no longer scale. Thus it is of interest to develop a fast heuristic for such instances.

A feasible solution  $\beta = (\beta_1, \dots, \beta_K)$  to problem (17) can be written as a vector of indices  $\mathbf{i} = (i_1, \dots, i_K) \in \{1, \dots, d\}^K$  and a vector of values  $\delta = (\delta_1, \dots, \delta_K) \in \mathbb{R}^K$ , such that for  $0 \leq k < K$ ,

$$(\beta_{k+1})_i = \begin{cases} (\beta_k)_i + \delta_k, & \text{if } i = i_k \\ (\beta_k)_i, & \text{if } i \neq i_k. \end{cases}$$

The vector of indices  $\mathbf{i}$  encodes which regression coefficients are modified at each step in the interpretable path, while the sequence of values  $\delta$  encodes the value of each modified regression coefficient. Thus problem (17) can be rewritten as

$$\min_{\mathbf{i}} \min_{\delta} C(\mathbf{i}, \delta), \quad \text{with } C(\mathbf{i}, \delta) := \sum_{k=1}^K c(\beta_0 + \sum_{j=1}^k \delta_j e_{i_j}), \quad (18)$$

where  $e_i$  designates the  $i$ -th unit vector. Notice that the inner minimization problem is an “easy” convex quadratic optimization problem, while the outer minimization problem is a “hard” combinatorial optimization problem. We propose the following local improvement heuristic for the outer problem: given a first sequence of indices  $\mathbf{i} = \mathbf{i}^0$ , we randomly sample one step  $\kappa$  in the interpretable path. Keeping all  $i_k$  constant for  $k \neq \kappa$ , we iterate through all  $d$  possible values of  $i_\kappa$  and obtain  $d$  candidate vectors  $\hat{\mathbf{i}}$ . For each candidate, we solve the inner minimization problem and keep the one with lowest cost. The method is described in full detail as Algorithm 1, in the more general case where we sample not one but  $q$  steps from the interpretable path.

---

**Algorithm 1** Local improvement heuristic. Inputs: regression cost function  $c(\cdot)$ ; starting vector of indices  $\mathbf{i}^0$ . Parameters:  $q \in \mathbb{N}$  controls the size of the neighborhood,  $T \in \mathbb{N}$  controls the number of iterations.

---

```

1: function LOCALIMPROVEMENT( $c(\cdot)$ ,  $\mathbf{i}^0$ ,  $q$ ,  $T$ )
2:   for  $1 \leq t \leq T$  do
3:      $\mathbf{i}^* \leftarrow \mathbf{i}^0$ 
4:      $\delta^* \leftarrow \arg \min_{\delta} C(\mathbf{i}^0, \delta)$ 
5:      $C^* \leftarrow C(\mathbf{i}^0, \delta^*)$ 
6:     Randomly select  $\mathcal{K} = \{\kappa_1, \dots, \kappa_q\} \subset \{1, \dots, K\}$  ▷ subset of cardinality  $q$ 
7:      $\hat{\mathbf{i}} \leftarrow \mathbf{i}^*$ 
8:      $\hat{\delta} \leftarrow \delta^*$ 
9:     for  $(f_1, \dots, f_q) \in \{1, \dots, d\}^q$  do
10:      for  $1 \leq p \leq q$  do
11:         $\hat{i}_{\kappa_p} = f_p$ 
12:         $\hat{\delta} \leftarrow \arg \min_{\delta} C(\hat{\mathbf{i}}, \delta)$ 
13:        if  $C(\hat{\mathbf{i}}, \hat{\delta}) < C^*$  then
14:           $C^* \leftarrow C(\hat{\mathbf{i}}, \hat{\delta})$ 
15:           $\mathbf{i}^* \leftarrow \hat{\mathbf{i}}$ 
16:           $\delta^* \leftarrow \hat{\delta}$ 
17:   return  $\mathbf{i}^*, \delta^*$ 

```

---

In order to empirically evaluate the local improvement heuristic, we run it with different batch sizes  $q$  on a small real dataset, with 100 rows and 6 features (after one-hot encoding of categorical features). The goal is to predict the perceived prestige (from a survey) of a job occupation given features about it, including education level, salary, etc.

Given this dataset, we first compute the optimal coordinate path of length  $K = 10$ . We then test our local improvement heuristic on the same dataset. Given the small size of the problem, in the complete formulation a provable global optimum is found by Gurobi in about 5 seconds. To be useful, we would like our local improvement heuristic to find a good solution significantly faster. We show convergence results of the heuristic for different values of the batch size parameter  $q$  in Table 2. For both batch sizes, the local improvement heuristic converges two orders of magnitude faster than Gurobi. With a batch size  $q = 2$ , the solution found is optimal.

### 6.2.1 Results

We now explore the results of the presented approach on a dataset of test scores in California from 1998-1999. Each data point represents a school, and the variable of interest is the average standardized test score of students from that school. All features are continuous and a full list is presented in Table 9a. Both the features and the target variables are centered and rescaled to have unit variance.

In our example, we assume that we already have a regression model available to predict the averaged test score: it was trained using only the percentage of students qualifying for a reduced-price lunch. This model has an MSE of 0.122 (compared to an optimal MSE of 0.095). We would like to update this model in an interpretable way given the availability of all features in the dataset. This corresponds to the problem of constructing an interpretable path, as before, with the simple modification that  $m_0$  no longer designates the regression model  $\mathbf{0}$ , but an arbitrary starting model (in particular, the one we have been provided).

The first thing we can do is explore the price of interpretability in this setting. We can use the method presented in Section 5.1 to compute find Pareto efficient interpretable models. The resulting price curve is shown in Figure 8.

| Method                 | Time (s) | Gap (%) |
|------------------------|----------|---------|
| Exact                  | 5.078    | 0.00    |
| Local imp. ( $q = 1$ ) | 0.004    | 0.02    |
| Local imp. ( $q = 2$ ) | 0.019    | 0.00    |

Table 2: Convergence time and optimality gap of local improvement heuristics for different batch sizes  $q$ .

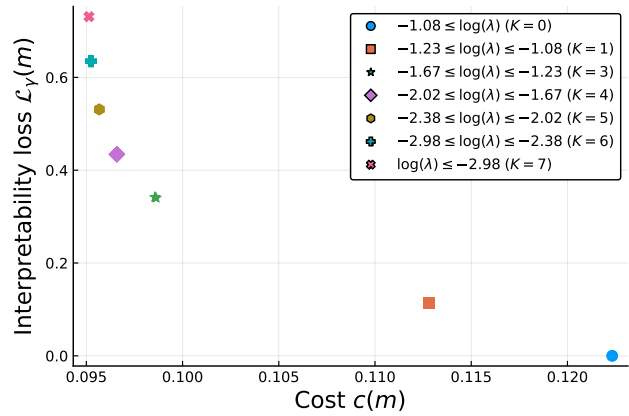


Figure 8: Pareto-efficient models from the perspective of interpretability and cost.

Given this price curve, we choose  $\log(\lambda) \approx -1.65$  because it yields an accurate final model while avoiding diminishing interpretability returns. This yields the new model (and associated interpretable path) shown in Figure 9b. This new model can be obtained from the old in just four steps. First we add the district average income with a positive coefficient, then we correct the coefficient for reduced-price lunch students to account for this new feature, and finally we add the percentage of English learners and the school’s per-student spending. The final model has an MSE of 0.097 which is near-optimal. When we compare this path to other methods (see Figure 9c) we see that our interpretable formulation allows us to find a good tradeoff between a greedy, “every step must improve” formulation, and a formulation that just sets the coefficients to their final values one by one.

## 7 Conclusions

In this paper, we have presented a simple optimization-based framework to model the interpretability of machine learning models. Our framework provides a new way to think about what interpretability means to users in different applications and quantify how this meaning affects the tradeoff with predictive accuracy. This framework is general, and each application can have its own modeling and optimization challenges, and could be an opportunity for further research.

## Acknowledgements

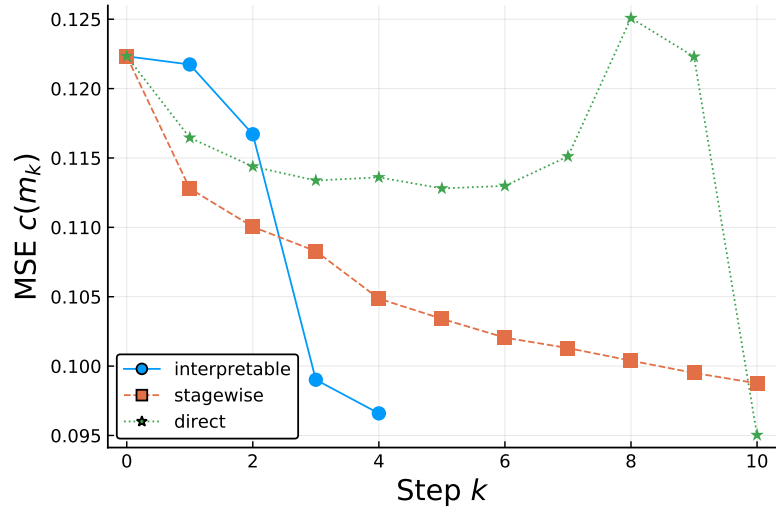
Research funded in part by ONR grant N00014-18-1-2122.

| Feature name | Description               |
|--------------|---------------------------|
| Enrollment   | Total enrollment          |
| Teachers     | Number of teachers        |
| CalWPct      | % receiving state aid     |
| MealPct      | % with subsidized lunch   |
| Computers    | Number of computers       |
| CompStu      | Computers per student     |
| ExpnStu      | Expenditure per student   |
| StuTeach     | Student-teacher ratio     |
| AvgInc       | Average income (district) |
| ELPct        | % English Learners        |

(a) Features of the test score dataset.

| Step | Feature      |             |              |             | MSE   |
|------|--------------|-------------|--------------|-------------|-------|
|      | MealPct      | AvgInc      | ELPct        | ExpnStu     |       |
| 0    | -0.87        | -           | -            | -           | 0.122 |
| 1    | -0.87        | <b>0.23</b> | -            | -           | 0.122 |
| 2    | <b>-0.59</b> | 0.23        | -            | -           | 0.117 |
| 3    | -0.59        | 0.23        | <b>-0.18</b> | -           | 0.099 |
| 4    | -0.59        | 0.23        | -0.18        | <b>0.07</b> | 0.097 |

(b) Path from old model to new model.



(c) Comparison between interpretable path and other approaches

Figure 9: Example of a Pareto-efficient interpretable path. On the left we see the benefits of each coefficient modification. On the right we compare the interpretable path with two other possible paths. The first is the forward stagewise path which greedily selects the best  $m_{k+1}$  given  $m_k$ . The second is a “direct” path, which adds the optimal least squares coefficients one by one. The direct method is only good when all the coefficients have been added, whereas the greedy approach is good at first but then does not converge. The interpretable path is willing to make some steps that do not improve the cost too much in preparation for very cost-improving steps.

## References

- [1] Hamsa Bastani, Osbert Bastani, and Carolyn Kim. Interpreting Predictive Models for Human-in-the-Loop Analytics. *arXiv preprint arXiv:1705.08504*, pages 1–45, 2018.
- [2] Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, 2017.
- [3] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.
- [4] Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, 40(2):210–217, 2017.
- [5] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.
- [6] Dimitris Bertsimas and Bart Van Parys. Sparse High-Dimensional Regression: Exact Scalable Algorithms and Phase Transitions. *Annals of Statistics*, to appear, 2019.
- [7] Leo Breiman. *Classification and regression trees*. New York: Routledge, 1984.
- [8] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [9] Leo Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.
- [10] Cristian Bucil, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 535, New York, New York, USA, 2006. ACM, ACM Press.
- [11] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic Transparency via Quantitative Input Influence. In *2016 IEEE Symposium on Security and Privacy*, 2016.
- [12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
- [13] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, nov 2016.

- [14] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, (MI):1–13, 2017.
- [15] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *Annals of Statistics*, 32(2):407–499, apr 2004.
- [16] Alex A. Freitas. Comprehensible classification models. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.
- [17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, NY, USA:, 2001.
- [18] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations : An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069*, 2018.
- [19] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". pages 1–9, 2016.
- [20] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.
- [21] Been Kim, Cynthia Rudin, and Julie Shah. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In *Neural Information Processing Systems (NIPS) 2014*, 2014.
- [22] I. Y. Kim and O. L. De Weck. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Structural and Multidisciplinary Optimization*, 29(2):149–158, 2005.
- [23] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2017.
- [24] Ali Koç and David P. Morton. Prioritization via Stochastic Optimization. *Management Science*, 61(3):586–603, 2014.
- [25] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: a joint framework for description and prediction. *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1:1675–1684, 2016.
- [26] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. *FAT/ML*, jul 2017.
- [27] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.
- [28] Guolong Lin and David Williamson. A general approach for incremental approximation and hierarchical clustering. *SIAM Journal Computing*, 39(8):3633–3669, 2010.
- [29] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*, 2016.
- [30] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible Models for Classification and Regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.
- [31] Sendhil Mullainathan and Ziad Obermeyer. Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–480, 2017.
- [32] Marco Tulio 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.
- [33] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, jun 2015.
- [34] Robert J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [35] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.
- [36] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.



## A Appendix

### A.1 Proof of Theorem 1

*Proof of part (a).* As  $c(\cdot)$  is bounded, we have  $c_{\max} \in \mathbb{R}$  such that  $0 < c(\cdot) \leq c_{\max}$ .

Let  $\mathbf{m}^+ \in \mathcal{P}(m^+)$  be a path of optimal length to the model  $m^+$ , i.e.,  $|\mathbf{m}^+| = \mathcal{L}_{\text{complexity}}(m^+)$ . Let  $\mathbf{m}^- \in \mathcal{P}(m^-)$  be any path leading to  $m^-$  (not necessarily of optimal length). By assumption, we have  $|\mathbf{m}^-| \geq |\mathbf{m}^+|$ , and by definition of model interpretability, we have  $\mathcal{L}_\gamma(m^+) \leq \mathcal{L}_\gamma(\mathbf{m}^+)$ . Therefore we obtain:

$$\mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(m^+) \geq \mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(\mathbf{m}^+) \quad (19)$$

$$= \sum_{k=1}^{|\mathbf{m}^-|} \gamma^k c(m_k^-) - \sum_{k=1}^{|\mathbf{m}^+|} \gamma^k c(m_k^+) \quad (20)$$

$$= \gamma^{|\mathbf{m}^+|} \left( \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} (c(m_k^-) - c(m_k^+)) + (c(m_{|\mathbf{m}^+|}^-) - c(m_{|\mathbf{m}^+|}^+)) + \sum_{k=|\mathbf{m}^+|+1}^{|\mathbf{m}^-|} \gamma^{k-|\mathbf{m}^+|} c(m_k^-) \right) \quad (21)$$

$$\geq \gamma^{|\mathbf{m}^+|} \left( -c_{\max} \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} + (c(m_{|\mathbf{m}^+|}^-) - c(m^+)) + \sum_{k=|\mathbf{m}^+|+1}^{|\mathbf{m}^-|} \gamma^{k-|\mathbf{m}^+|} c(m_k^-) \right), \quad (22)$$

where (20) follows from the definition of model interpretability, (21) is just a development of the previous equation, and (22) just bounds the first sum and uses  $m_{|\mathbf{m}^+|}^+ = m^+$  for the middle term.

If  $\mathcal{L}_{\text{complexity}}(m^+) < \mathcal{L}_{\text{complexity}}(m^-)$ , we have  $|\mathbf{m}^+| < |\mathbf{m}^-|$ , and therefore the last sum in (22) is not empty and for  $\gamma \geq 1$  we can bound it:

$$\sum_{k=|\mathbf{m}^+|+1}^{|\mathbf{m}^-|} \gamma^{k-|\mathbf{m}^+|} c(m_k^-) \geq \gamma^{|\mathbf{m}^-|-|\mathbf{m}^+|} c(m_{|\mathbf{m}^-|}^-) \geq \gamma c(m^-). \quad (23)$$

Therefore, for  $\gamma \geq 1$  we have:

$$\mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(m^+) \geq \gamma^{|\mathbf{m}^+|} \left( \gamma c(m^-) - c(m^+) - c_{\max} \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} \right). \quad (24)$$

This bound is valid for all the path  $\mathbf{m}^-$  leading to  $m^-$ , in particular the one with optimal interpretability loss, therefore we have (for  $\gamma \geq 1$ ):

$$\mathcal{L}_\gamma(m^-) - \mathcal{L}_\gamma(m^+) \geq \gamma^{|\mathbf{m}^+|} \left( \gamma c(m^-) - c(m^+) - c_{\max} \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} \right). \quad (25)$$

which implies (as  $c(m^-) > 0$ ):

$$\lim_{\gamma \rightarrow +\infty} \mathcal{L}_\gamma(m^-) - \mathcal{L}_\gamma(m^+) = +\infty \quad (26)$$

We now look at the case  $\mathcal{L}_{\text{complexity}}(m^+) = \mathcal{L}_{\text{complexity}}(m^-)$  and  $c(m^+) < c(m^-)$ . For  $\gamma \geq 1$ , we can easily bound parts of equation (22):

$$c(m_{|\mathbf{m}^+|}^-) + \sum_{k=|\mathbf{m}^+|+1}^{|\mathbf{m}^-|} \gamma^{k-|\mathbf{m}^+|} c(m_k^-) \geq \gamma^{|\mathbf{m}^-|-|\mathbf{m}^+|} c(m_{|\mathbf{m}^-|}^-) \geq c(m^-). \quad (27)$$

Putting it back into (22), we obtain (for  $\gamma \geq 1$ )

$$\mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(m^+) \geq \gamma^{|\mathbf{m}^+|} \left( (c(m^-) - c(m^+)) - c_{\max} \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} \right). \quad (28)$$

This bound is independent of the path  $\mathbf{m}^-$  leading to  $m^-$ , therefore we have

$$\mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(\mathbf{m}^+) \geq \gamma^{|\mathbf{m}^+|} \left( (c(\mathbf{m}^-) - c(\mathbf{m}^+)) - c_{max} \sum_{k=1}^{|\mathbf{m}^+|-1} \frac{1}{\gamma^{|\mathbf{m}^+|-k}} \right) \xrightarrow{\gamma \rightarrow +\infty} +\infty, \quad (29)$$

which ends the proof of part (a) of Theorem 1.  $\square$

*Proof of part (b).* Consider two paths  $\mathbf{m}^+, \mathbf{m}^- \in \mathcal{P}$ , such that  $\mathbf{c}(\mathbf{m}^+) \preceq \mathbf{c}(\mathbf{m}^-)$ . By definition of the lexicographic order, either the two paths are the same (in that case the theorem is trivial), or there exist  $K \geq 1$  such that:

$$\begin{cases} \mathbf{c}(\mathbf{m}^+)_k = \mathbf{c}(\mathbf{m}^-)_k & \forall k < K \\ \mathbf{c}(\mathbf{m}^+)_K < \mathbf{c}(\mathbf{m}^-)_K. \end{cases}$$

We have:

$$\mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(\mathbf{m}^+) = \sum_{k=1}^{|\mathbf{m}^-|} \gamma^k c(\mathbf{m}^-)_k - \sum_{k=1}^{|\mathbf{m}^+|} \gamma^k c(\mathbf{m}^+)_k \quad (30)$$

$$= \sum_{k=1}^{\infty} \gamma^k (c(\mathbf{m}^-)_k - c(\mathbf{m}^+)_k) \quad (31)$$

$$= \sum_{k=1}^{K-1} \gamma^k (c(\mathbf{m}^-)_k - c(\mathbf{m}^+)_k) + \gamma^K (c(\mathbf{m}^-)_K - c(\mathbf{m}^+)_K) + \sum_{k=K+1}^{\infty} \gamma^k (c(\mathbf{m}^-)_k - c(\mathbf{m}^+)_k) \quad (32)$$

$$= \gamma^K \left( c(\mathbf{m}^-)_K - c(\mathbf{m}^+)_K + \sum_{k=K+1}^{\infty} \gamma^{k-K} (c(\mathbf{m}^-)_k - c(\mathbf{m}^+)_k) \right), \quad (33)$$

where (31) just applies the definition of the sequence  $\mathbf{c}$ , and (33) uses  $\mathbf{c}(\mathbf{m}^+)_k = \mathbf{c}(\mathbf{m}^-)_k \quad \forall k < K$ .

The term inside the parenthesis in (33) converges to  $c(\mathbf{m}^-)_K - c(\mathbf{m}^+)_K > 0$  when  $\gamma \rightarrow 0$ , as the paths are finite. Therefore

$$\lim_{\gamma \rightarrow 0} \mathcal{L}_\gamma(\mathbf{m}^-) - \mathcal{L}_\gamma(\mathbf{m}^+) \geq 0, \quad (34)$$

which proves (8). The very end of the theorem is an immediate consequence.  $\square$

## A.2 Proof of Proposition 1

*Proof.* First, a solution of

$$\min_{\mathbf{m} \in \mathcal{M}} (c(\mathbf{m}) + \lambda \mathcal{L}_\alpha(\mathbf{m}))$$

is Pareto optimal between the cost  $c(\cdot)$  and the interpretability  $\mathcal{L}_\alpha(\cdot)$  as it corresponds to the minimization of a weighted sum of the objectives. Furthermore, we can write

$$\begin{aligned} \min_{\mathbf{m} \in \mathcal{M}} (c(\mathbf{m}) + \lambda \mathcal{L}_\alpha(\mathbf{m})) &= \min_{\mathbf{m} \in \mathcal{M}} \left( c(\mathbf{m}) + \lambda \min_{\mathbf{m} \in \mathcal{P}(\mathbf{m})} \mathcal{L}_\alpha(\mathbf{m}) \right) = \min_{\mathbf{m} \in \mathcal{M}, \mathbf{m} \in \mathcal{P}(\mathbf{m})} (c(\mathbf{m}) + \lambda \mathcal{L}_\alpha(\mathbf{m})) \\ &= \min_{\mathbf{m} \in \mathcal{M}, K \geq 0, \mathbf{m} \in \mathcal{P}_K(\mathbf{m})} \left( c(\mathbf{m}_K) + \lambda \sum_{k=1}^K \alpha_k c(\mathbf{m}_k) \right) \\ &= \min_{K \geq 0, \mathbf{m} \in \mathcal{P}_K} \left( c(\mathbf{m}_K) + \lambda \sum_{k=1}^K \alpha_k c(\mathbf{m}_k) \right). \end{aligned}$$

$\square$

### A.3 Proof of Proposition 2

*Proof.* For any  $K \geq 0$  and  $\lambda > 0$ , define the optimal objective

$$z_\lambda(K) = \min_{\mathbf{m} \in \mathcal{P}_K} c(m_K) + \lambda \sum_{k=1}^K \gamma^k c(m_k).$$

Because  $c(\cdot)$  is bounded below by  $c_{\min}$ , we can write

$$z_\lambda(K) \geq c_{\min} + \lambda \sum_{k=1}^K \gamma^k c_{\min} \geq \lambda c_{\min} \sum_{k=1}^K \gamma^k. \quad (35)$$

By definition  $z_\lambda(0)$  is the cost of the empty model, so by the boundedness of  $c(\cdot)$ , we have  $z_\lambda(0) \leq c_{\max}$ . Consider first the case when  $\gamma = 1$ . Then (35) simplifies to

$$z_\lambda(K) \geq \lambda K c_{\min}.$$

Setting  $K \geq K_{\max} := c_{\max}/(\lambda c_{\min})$  yields  $z_\lambda(K) \geq c_{\max} \geq z_\lambda(0)$  and so the interpretable path of length 0 has a better objective than any path of length at least  $K_{\max}$ . Now consider  $\gamma > 1$ . In this case, (35) simplifies to

$$z_\lambda(K) \geq \lambda c_{\min} \gamma \frac{1 - \gamma^K}{1 - \gamma}.$$

Defining  $K_{\max}$  as in (13), we again see that the interpretable path of length 0 has a better objective than any path of length at least  $K_{\max}$ , which completes the proof.  $\square$