# Probabilistic Explanations for Linear Models*

Bernardo Subercaseaux ⓣ

bersub@cmu.edu

Carnegie Mellon University

Marcelo Arenas ⓣ

marenas@ing.puc.cl
PUC Chile
IMFD Chile
RelationalAI

Kuldeep S. Meel ⓣ

meel@cs.toronto.edu
Georgia Institute of Technology
University of Toronto

December 19, 2024

Abstract

Formal XAI is an emerging field that focuses on providing explanations with mathematical guarantees for the decisions made by machine learning models. A significant amount of work in this area is centered on the computation of "sufficient reasons". Given a model $\mathcal{M}$ and an input instance x, a sufficient reason for the decision $\mathcal{M}(x)$ is a subset S of the features of x such that for any instance z that has the same values as x for every feature in S, it holds that $\mathcal{M}(x) = \mathcal{M}(z)$. Intuitively, this means that the features in S are sufficient to fully justify the classification of x by $\mathcal{M}$. For sufficient reasons to be useful in practice, they should be as small as possible, and a natural way to reduce the size of sufficient reasons is to consider a probabilistic relaxation; the probability of $\mathcal{M}(x) = \mathcal{M}(z)$ must be at least some value $\delta \in (0, 1]$, for a random instance z that coincides with x on the features in S. Computing small δ-sufficient reasons (δ-SRs) is known to be a theoretically hard problem; even over decision trees — traditionally deemed simple and interpretable models — strong inapproximability results make the efficient computation of small δ-SRs unlikely. We propose the notion of (δ, ϵ)-SR, a simple relaxation of δ-SRs, and show that this kind of explanations can be computed efficiently over linear models.

1 Introduction

Explaining the decisions of Machine Learning classifiers is a fundamental problem in XAI (Explainable AI), and doing so with formal mathematical guarantees on the quality, size, and semantics of the explanations is in turn the core of *Formal XAI* (Marques-Silva and Ignatiev, 2022). Within formal XAI, one of the most studied kinds of explanations is that of *sufficient reasons* (Darwiche and Hirth, 2020), which aim to explain a decision $\mathcal{M}(x) = 1$ by presenting a subset S of the features of the input x that implies $\mathcal{M}(z) = 1$ for any z that agrees with x on S. In the language of theoretical computer science, these correspond to *certificates* for $\mathcal{M}(x)$.

Example 1. *Consider a binary classifier $\mathcal{M}$ defined as*

$$\mathcal{M}(x) = (x_1 \vee \overline{x_3}) \wedge (x_2 \vee \overline{x_1}) \wedge (x_4 \vee x_3),$$

---

*The authors opted for randomized author ordering, denoted by ⓣ, instead of alphabetical. A verifiable record of the randomization is available at https://www.aeaweb.org/journals/policies/random-author-order/search

and the input instance $x = (1, 1, 0, 1)$. We can say that $\mathcal{M}(x)$ "because" $x_1 = 1, x_2 = 1$, and $x_4 = 1$, as they are sufficient to determine the value of $\mathcal{M}(x)$ regardless of x_3.

Let us start formalizing the framework for our work. First, we consider binary boolean models $\mathcal{M}: \{0, 1\}^d \rightarrow \{0, 1\}$. Despite our domain being binary, we will need a third value, $\perp$, to denote *"unknown"* values. For example, we may represent a person who *does* have a car, *does not* have a house, and for whom we do not know if they have a pet or not, as $(1, 0, \perp)$. We say elements of $\{0, 1, \perp\}^d$ are *partial instances*, while elements of $\{0, 1\}^d$ are simply *instances*. To illustrate, in Example 1 we used the partial instance $y = (1, 1, \perp, 1)$ to explain $\mathcal{M}(x) = 1$. We use the notation $y \sqsubseteq x$ to denote that the (partial) instance x *"fills in"* values of the partial instance y; more formally, we use $y \sqsubseteq x$ to mean that $y_i = \perp \vee y_i = x_i$ for every $i \in [d]$. Finally, for any partial instance y we denote by $\text{Comp}(y)$ the set of instances x such that $y \sqsubseteq x$, thinking of $\text{Comp}(y)$ as the set of *completions* of y. One can define sufficient reasons as follows with this notation.

Definition 1 (Sufficient Reason (Darwiche and Hirth, 2020))**.** *We say y is a* sufficient reason *for x if for any completion $z \in \text{Comp}(y)$ it holds that $\mathcal{M}(x) = \mathcal{M}(z)$.*

A crucial factor for the helpfulness of sufficient reasons as explanations is their size; even though x is always a sufficient reason for its own classification, we long for explanations that are much smaller than x itself. Miller (1956), for instance, goes on to say that explanations consisting of more than 9 features are probably too large for human stakeholders. In general, empirical research suggests that explanations ought to be small (Narayanan et al., 2018; Lage et al., 2019). There are several ways of formalizing the succinctness we desire for sufficient reasons:

- **(Minimum Size)** For a sufficient reason y, we define its *explanation size* $|y|_e$ as the number of defined features in y, or equivalently, $|y|_e := d - |y|_\perp$, where $|y|_\perp$ is the number of features of y taking $\perp$. See e.g., Barceló et al. (2020).[1]

- **(Subset minimality)** We say a sufficient reason y for a pair $(\mathcal{M}, x)$ is *minimal* if there is no other sufficient reason y' for $(\mathcal{M}, x)$ such that $y' \subsetneq y$. In fact, the original definition of sufficient reasons of Darwiche and Hirth (2020) includes minimality as a requirement, and so is the case under the *"abductive explanation"* naming (Ignatiev et al., 2021).

- **(Relative to average explanation)** Blanc et al. (2021) compute explanations that are small relative to the *"certificate complexity"* of the classifier $\mathcal{M}$, meaning the average size of the minimum sufficient reason where the average is taken over all possible instances x.

Nevertheless, there is a path toward even smaller explanations: *probabilistic* sufficient reasons (Wäldchen et al., 2021; Izza et al., 2023a). As will be shown in Example 2., and is noted as a remark by Blanc et al. (2021), these can be arbitrarily smaller than minimum size sufficient reasons.

2 Probabilistic Sufficient Reasons

The main idea of probabilistic sufficient reasons is to relax the condition *"all completions of the explanation y have the same class as x"* to *"a random completion of y has the same class as x with high probability"*.

---

[1]When talking about a partial instance y, we will use the "size" of y to mean $|y|_e$.

Let us use notation $z \sim \mathbf{U}(y)$ to denote that z is a completion of y drawn uniformly at random. With this notation we can define δ-sufficient reasons:[2]

Definition 2 (Wäldchen et al. (2021)). *For any $\delta \in [0, 1]$, a δ-sufficient reason (δ-SR) for an instance x, is a partial instance $y \subseteq x$ such that*

$$\Pr_{z \sim \mathbf{U}(y)} \big[\mathcal{M}(z) = \mathcal{M}(x) \big] \geqslant \delta.$$

Naturally, a minimum δ-SR is a δ-SR of minimum size. Note immediately that Definition 2 and Definition 1 coincide when $\delta = 1$.

2.1 The size of δ-SRs

Interestingly, even a 0.999999-SR can be arbitrarily smaller, in terms of defined features, than the smallest sufficient reason (i.e., 1-SR) for a pair $(\mathcal{M}, x)$, even when $\mathcal{M}$ is a linear model, as we will illustrate in Example 2. Before providing the example, let us define linear models.

Definition 3. *A (binary) linear model $\mathcal{L}$ of dimension d is a pair (w, t), where $w \in \mathbb{Q}^d$ and $t \in \mathbb{Q}$. Its classification over an instance x is defined simply as*

$$\mathcal{L}(x) = \begin{cases} 1 & \text{if } x \cdot w \geqslant t \\ 0 & \text{otherwise.} \end{cases}$$

Example 2. *Consider a linear model $\mathcal{L}$ of dimension $d = 1000$ with parameters $t = 1250$ and*

$$w = (1000, 1, 1, 1, 1, \ldots, 1).$$

Let the instance x be $(1, 1, 1, 1, 1, \ldots, 1)$, so that clearly $\mathcal{L}(x) = 1$. One can easily see that any 1-SR for x under $\mathcal{L}$ has size 251, as it must include the first feature and any 250 other features. However, if we consider $y = (1, \perp, \perp, \perp, \ldots, \perp)$, then a simple application of the Chernoff-Hoeffding concentration bound (in the appendix for the completeness) gives that

$$\Pr_{z \sim \mathbf{U}(y)} \big[\mathcal{L}(z) = 1 \big] \geqslant 0.999999.$$

This suggests that we might say $\mathcal{L}(x) = 1$ "because" $x_1 = 1$; formally, y is a 0.999999-SR, and 251 times smaller than any 1-SR for $\mathcal{L}(x)$.

3 Approximating δ-Sufficient Reasons

Unfortunately, computing small δ-SRs is computationally challenging, even when attempting to find approximate solutions. Let us contextualize our main result by summarizing first what is known about the complexity of computing δ-SRs and their deterministic predecessors, 1-SRs.

---

[2]Also known as δ-relevant sets Izza et al. (2021); Wäldchen et al. (2021).

Barceló et al. (2020) showed that computing a minimum 1-SR is Σ_2^p-hard for neural networks, NP-hard for decision trees, and polynomial-time solvable for linear models. Then, Wäldchen et al. (2021, Theorem 2.4) showed that computing minimum δ-SRs for neural networks is hard for NP$^{\text{PP}}$, and Arenas et al. (2022) proved that even for the restricted class of decision trees, which are usually considered interpretable, minimum δ-SRs cannot be computed in polynomial time unless P = NP (and neither can subset-minimal δ-SRs for $\delta < 1$, in contrast to the $\delta = 1$ setting which is in P (Izza et al., 2020; Subercaseaux, 2020)). For linear models, even computing the value

$$\Pr_{z \sim \mathbf{U}(y)} \left[\mathcal{L}(z) = \mathcal{L}(x) \right]$$

exactly is #P-hard (Barceló et al., 2020), from where the following is easy to show.[3][4]

Proposition 1. *Given a linear model $\mathcal{L}$, an instance x, and $\delta \in [0,1]$, the size of the smallest δ-SR for $(\mathcal{L}, x)$ cannot be computed in polynomial time unless* FP = #P.

Furthermore, the situation does not improve if we aim to efficiently approximate the value $\text{MIN}(\mathcal{M}, x, \delta)$. Wäldchen et al. (2021, Theorem 2.5) studied general classifiers (e.g., neural networks) and showed that no algorithm can achieve an approximation factor of $d^{1-\alpha}$ for this problem, where d is the dimension of the classifier and $\alpha > 0$, unless P = NP. Kozachinskiy (2023) proved that this approximation task is also hard for decision trees.

However, these hardness results do not preclude the existence of efficient algorithms for computing or approximating δ-SR for linear models. Hence, the goal of this section is to explore these questions for such models, given their practical importance.

3.1 A Simple Relaxation: (δ, ε)-min-SR

In light of the hardness results for δ-SRs, it is natural to consider a further relaxation that would allow for tractability. Consider for instance a customer of a bank who wants a 0.95-SR for why their application for a loan was rejected. Such an explanation would consist of a small number of features of their application profile that are relevant to the decision since 95% of applicants with such a profile would also get rejected. We expect that, in such a scenario, the user would not particularly care if the explanation she obtains holds for 95% of potential applicants or for 94.9997% of them. In other words, the value of δ is chosen in a trade-off between the size of the explanation and the desired level of confidence or "explanation power". We posit that in such a trade-off, the user is more sensitive to increases in the explanation size than they are to a minor perturbation in δ, the probability guarantee. This motivates the following definition:

Definition 4 ((δ, ε)-min-SR). *Given a model $\mathcal{M}$, an instance x, and values $\delta, \varepsilon \in (0,1)$, we say a partial instance y of size is a (δ, ε)-min-SR if there exists a value $\delta^\star \in [\delta - \varepsilon, \delta + \varepsilon]$ such that y is a minimum $\delta^\star$-SR for x under $\mathcal{M}$.*

Note that, even though the guarantee of a (δ, ε)-min-SR is symmetric around δ, our definition is such that the ability of efficiently computing (δ, ε)-min-SRs is enough for the following two tasks:

---

[3]Izza et al. (2023b) already made a more general observation of this form, but did not provide a hardness result for linear models.

[4]All proofs that are not presented in the main text can be found in the appendix.

1. A user wants an explanation as small as possible and of probability "close" to δ. Then, by computing a $(\delta - \varepsilon/2, \varepsilon/2)$-min-SR, they obtain an explanation whose probability guarantee is at most ε away from δ, and is no larger in size than the minimum δ-SR.

2. The owner of the model wants to offer a δ-SR that is as small as possible to a customer, and they want to be strict on the δ part, since offering a $(\delta - \varepsilon)$-SR would be misleading and could lead to legal issues. Then, by computing a $(\delta + \varepsilon/2, \varepsilon/2)$-min-SR, they can guarantee that the explanation is at least δ-SR, while still being likely much smaller than a minimum 1-SR.

The inapproximability result of Kozachinskiy (2023) can be translated to the (δ, ε)-min-SR problem as follows:

Theorem 1 (Kozachinskiy (2023), Theorem 1). *Unless* SAT *can be solved in quasi-polynomial time, one cannot compute a (δ, ε)-min-SR for decision trees in polynomial time, and furthermore, any polynomial-time algorithm that guarantees to provide a δ'-SR for some $\delta' \in [\delta - \varepsilon, \delta + \varepsilon]$ will produce explanations that are up to $\Omega(d^{1-\alpha})$ times larger than any (δ, ε)-min-SR, for any $\alpha > 0$.*

Note that this hardness result for decision trees implies in turn hardness for neural networks by using standard compilation techniques (Barceló et al., 2020). Our main result is that, for linear models, we can efficiently compute (δ, ε)-min-SRs, making them the first class of models for which we have such a positive result. To state our runtime more cleanly, we use the standard notation $\widetilde{O}(f)$ to mean $O(f \cdot \log(f)^c)$ for some constant $c \in \mathbb{R}$.

Theorem 2. *Given a linear model $\mathcal{L}$ and an input x, we can compute a (δ, ε)-min-SR successfully with probability $1 - \gamma$ in time $\widetilde{O}\left(\frac{d}{\varepsilon^2 \gamma^2}\right)$; that is, polynomial in d, $1/\varepsilon$, and $1/\gamma$.*

We remark that previous approaches for computing approximate probabilistic explanations lacked theoretical guarantees on the size of the explanations produced (Izza et al., 2023b, 2021, 2024).

In order to prove Theorem 2 we will need two main ideas: (i) the fact that we can estimate the probabilities of models accepting a partial instance through Monte Carlo sampling[5], and (ii) that under the uniform distribution it is easy to decide which features ought to be part of small explanations.

3.2 Estimating the Probability of Acceptance

The hardness of computing $\Pr_{z \sim \mathbf{U}(y)}[\mathcal{M}(z) = 1]$ is about computing it to arbitrarily high precision, i.e., with an additive error within $O(2^{-d})$. However, computing a less precise estimation of $\Pr_{z \sim \mathbf{U}(y)}[\mathcal{M}(z) = 1]$ is simple, as the next fact (which is a direct consequence of Hoeffding (1963)'s inequality) states.

Fact 1. *Let f be an arbitrary boolean function on n variables. Let M be any positive integer, and let $x_1, \ldots, x_M$ be M uniformly random samples from $\{0, 1\}^n$. Then*

$$\widehat{\mu}(M) := \frac{\sum_{i=1}^{M}[f(x_i) = 1]}{M}$$

---

[5]This idea is already used in the work of Izza et al. (2024)

is an unbiased estimator for

$$\mu := \Pr_{x \in \{0,1\}^n}[f(x) = 1],$$

and

$$\Pr[|\hat{\mu}(M) - \mu| \leqslant t] \geqslant 1 - 2\exp(-2t^2 M),$$

which is at least $1 - \gamma$ *for* $M = \frac{1}{2t^2}\log(2/\gamma)$.

As a consequence of the previous idea, although a minimum δ-SR might be hard to compute, this crucially depends on the value of δ. In order to deal with this, our algorithm will sample a value $\delta^\star$ uniformly at random from $[\delta - \varepsilon, \delta + \varepsilon]$, and then compute a minimum $\delta^\star$-SR. Intuitively, the idea is that as $\delta^\star$ is chosen at random, it will be unlikely that a value that makes the computation hard is chosen.

Before proving Theorem 2, we need to prove a lemma concerning the easiness of selecting the features of the desired explanation.

3.3 Feature Selection

Even if we were granted an oracle computing the probabilities $\Pr_{z \in \mathbf{D}(y)}[\mathcal{M}(z) = 1]$, that would not be necessarily enough to efficiently compute a minimum δ-SR. Indeed, for decision trees, the counting problem can be easily solved in polynomial time (Barceló et al., 2020), and yet the computation of δ-SRs of minimum size is hard, even to approximate (Arenas et al., 2022; Kozachinskiy, 2023). Intuitively, the problem for decision trees is that, even if we were told that the minimum δ-SR has exactly k features, it is not obvious how to search for it better than enumerating all $\binom{d}{k}$ subsets. The case of linear models, however, is different, at least under the uniform distribution. In this case, every feature i that is not part of the explanation will take value 0 or 1 independently with probability $1/2$, and contribute to the classification according to its weight w_i. In other words, we can sort the features according to their weights (with some care about signs), and select them greedily to build a small δ-SR. A proof for the deterministic case ($\delta = 1$) was already given in Barceló et al. (2020) and sketched earlier on by Marques-Silva et al. (2020a).

Definition 5. *Given a linear model* $\mathcal{L} = (w, t)$, *and an instance* x, *both having dimension* d, *we define the* score *of feature* $i \in [d]$ *as*

$$s_i := w_i \cdot (2x_i - 1) \cdot (2\mathcal{L}(x) - 1).$$

In other words, the sign of s_i is $+1$ if the feature is "helping" the classification, and -1 if it is "hurting" it. The magnitude of s_i is proportional to the weight of the feature i. Changing the value of feature i in an instance x would decrease $w \cdot x$ by s_i if $\mathcal{L}(x) = 1$, and increase it by s_i if $\mathcal{L}(x) = 0$. For the uniform distribution (or more generally, any distribution in which all features are Bernoulli variables with the same parameter), we can prove the following lemma that basically states that, for linear models it is good to choose features greedily according to their score.

Lemma 1. *Given a linear model* $\mathcal{L}$, *and an instance* x, *if* $y^{(0)}, \ldots, y^{(d)}$ *are the partial instances of* x *such that* $y^{(k)} \subseteq x$ *is defined only in the top* k *features of maximum score, then*

$$\Pr_{z \sim \mathbf{U}(y^{(k+1)})}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \Pr_{z \sim \mathbf{U}(y^{(k)})}[\mathcal{L}(z) = \mathcal{L}(x)]$$

for all $k \in \{0, \ldots, d-1\}$, and naturally,

$$\Pr_{z \sim \mathbf{U}(y^{(d)})} [\mathcal{L}(z) = \mathcal{L}(x)] = 1.$$

Moreover, $\text{MIN}(\mathcal{L}, x, \delta) = k$ if and only if $y^{(k)}$ is a δ-SR for x, and either $k = 0$ or $y^{(k-1)}$ is not a δ-SR for x.

Even though a proof of Lemma 1 is presented in the appendix, let us provide a self-contained example that should convince a reader of the veracity of the lemma.

Example 3. *Consider an instance $x = (1, 0, 0, 1, 1)$ and the linear model $\mathcal{L}$ be defined by*

$$w = (5, 1, -3, 2, -1) \quad ; \quad t = 5.$$

It is easy to check that $w \cdot x = 6$, and thus $\mathcal{L}(x) = 1$. The feature scores, according to Definition 5, are:

$$s_1 = 5, \ s_2 = -1, \ s_3 = 3, \ s_4 = 2, \ s_5 = -1.$$

For the first part, the main idea is that a positive score s_i means that the feature is helping the classification (i.e., adding it to a partial instance does not decrease its probability guarantee), while a negative score means that the feature is hurting the classification (i.e., adding it to a partial instance does not increase its probability guarantee). Because the partial instances

$$y^{(0)} \subseteq y^{(1)} \subseteq \cdots y^{(d)}$$

are obtained by adding a single feature at a time, and thus features are added in decreasing order of their scores, then this procedure will have two phases: (i) First, it will add features with a positive score, which raise or maintain the probability of the classification being the same as x, as the lemma says, and then (ii) it will start adding features with a negative score, which would seem to contradict the lemma, but it turns out that at that point the partial instance $y^{(k)}$ would have probability guarantee 1; this is because $y^{(d)} = x$, which trivially has probability guarantee 1. Table 2 presents the probabilities associated to the partial instances $y^{(0)}, \ldots, y^{(d)}$.

For the second part, consider the partial instances $y^\star = (\bot, 0, 0, 1, 1)$ and $y^\dagger = (1, \bot, 0, 1, 1)$. The instance x is a completion of both $y^\star$ and $y^\dagger$, but $y^\star$ also has completion

$$x^\star = (0, 0, 0, 1, 1),$$

whereas $y^\dagger$ has also completion

$$x^\dagger = (1, 1, 0, 1, 1).$$

Note that $w \cdot x^\star = 1 = w \cdot x - s_1$, whereas $w \cdot x^\dagger = 6 = w \cdot x - s_2$. Intuitively, this means that it is better to keep feature 1 as part of the explanation, but not feature 2. If we want an explanation with only two features, we should choose feature 1 and feature 3, as they have the highest scores. Indeed, Table 1 presents the probabilities to all possible explanations of size 2.

With Lemma 1 in hand, we can proceed to prove Theorem 2.

Table 1: Table of probabilities associated to Example 3.

Partial instance	Features included	Probability
$(1, \perp, 0, \perp, \perp)$	$\{1, 3\}$	$7/8$
$(1, \perp, \perp, 1, \perp)$	$\{1, 4\}$	$5/8$
$(\perp, \perp, 0, 1, \perp)$	$\{3, 4\}$	$1/2$
$(\perp, \perp, 0, \perp, 1)$	$\{3, 5\}$	$3/8$
$(\perp, 0, 0, \perp, \perp)$	$\{2, 3\}$	$3/8$
$(1, \perp, \perp, \perp, 1)$	$\{1, 5\}$	$3/8$
$(1, 0, \perp, \perp, \perp)$	$\{1, 2\}$	$3/8$
$(\perp, \perp, \perp, 1, 1)$	$\{4, 5\}$	$1/4$
$(\perp, 0, \perp, 1, \perp)$	$\{2, 4\}$	$1/4$
$(\perp, 0, \perp, \perp, 1)$	$\{2, 5\}$	$1/8$

Table 2: Table of probabilities associated to Example 3. The last column denotes the score of the feature added to the partial instance in that row with respect to the previous row.

Partial instance		Features	Probability	Score
$\boldsymbol{y}^{(0)}$	$(\perp, \perp, \perp, \perp, \perp)$		$1/4$	-
$\boldsymbol{y}^{(1)}$	$(1, \perp, \perp, \perp, \perp)$		$1/2$	5
$\boldsymbol{y}^{(2)}$	$(1, \perp, 0, \perp, \perp)$		$7/8$	3
$\boldsymbol{y}^{(3)}$	$(1, \perp, 0, 1, \perp)$		$1/1$	2
$\boldsymbol{y}^{(4)}$	$(1, 0, 0, 1, \perp)$		$1/1$	-1
$\boldsymbol{y}^{(5)}$	$(1, 0, 0, 1, 1)$		$1/1$	-1

Algorithm 1 LinearMonteCarloExplainer

Input: A linear model $\mathcal{L}$, an instance x, and $\delta \in (0, 1)$.
Parameters: $\varepsilon \in (0, 1)$, $\gamma \in (0, 1)$.
Output: A value $\delta^\star \in [\delta - \varepsilon, \delta + \varepsilon]$ together with a minimum $\delta^\star$-SR explanation for x.

1: $\delta^\star \leftarrow$ uniformly random sample from $[\delta - \varepsilon, \delta + \varepsilon]$
2: **for** $i \in \{1, \ldots, d\}$ **do**
3:    $s_i \leftarrow w_i \cdot (2x_i - 1) \cdot (2\mathcal{L}(x) - 1)$
4: **end for**
5: **for** $k \in \{0, 1 \ldots, d\}$ **do**
6:    let $y^{(k)} \subseteq x$ be the partial instance defined only in the top k features with maximum score s_i.
7: **end for**
8: $M \leftarrow (\log^2 d)/(2\varepsilon^2 \gamma^2) \log(2 \log d / \gamma)$
9: LB $\leftarrow 0$, UB $\leftarrow d$, and STEPS $\leftarrow 0$
10: **while** LB $\neq$ UB and STEPS $\leqslant \log d$ **do**
11:    STEPS $\leftarrow$ STEPS $+ 1$
12:    $m \leftarrow (\text{LB} + \text{UB})/2$
13:    $\widehat{v}_m \leftarrow$ MONTECARLOESTIMATION$(\mathcal{L}, y^{(m)}, x, M)$
14:    **if** $\widehat{v}_m \geqslant \delta^\star$ **then**
15:      UB $\leftarrow m$
16:    **else**
17:      LB $\leftarrow (m + 1)$
18:    **end if**
19: **end while**
20: $k^\star \leftarrow$ LB (or equivalently, UB)
21: **return** $(\delta^\star, y^{(k^\star)})$

Proof of Theorem 2. We use Algorithm 1. Let us define the partial instances $y^{(0)}, y^{(1)} \ldots, y^{(d)}$ so that $y^{(k)} \subseteq x$ is the partial instance defined only in the k features with maximum score (line 6). We then define a sequence of values v_k as

$$v_k := \Pr_{z \sim \mathbf{U}(y^{(k)})} [\mathcal{L}(z) = \mathcal{L}(x)],$$

and note that due to Lemma 1, the sequence $v_0, \ldots, v_d$ is non-decreasing. Let $M = \frac{\log^2 d}{2\varepsilon^2 \gamma^2} \log(2 \log d / \gamma)$, as in line 8, and let us define random variables $\widetilde{v}_k$ as follows: if Algorithm 1 enters line 13 with $m = k$, then $\widetilde{v}_k$ is the output of Algorithm 2 (i.e., $\widehat{v}_k(M)$), and otherwise $\widetilde{v}_k = v_k$. We use binary search (lines 10-19), to find $k^\star$, the smallest k such that

$$\widetilde{v}_k \geqslant \delta^\star,$$

and our goal is to show that with good probability $k^\star$ is also the smallest k such that $v_k \geqslant \delta^\star$, which would imply the correctness of the algorithm by Lemma 1. Note, however, that even though the sequence $v_0, \ldots, v_d$ is non-decreasing (Lemma 1), the estimated values $\widehat{v}_k$ are not necessarily so.

Algorithm 2 MonteCarloEstimation

Input: A linear model $\mathcal{L}$, a partial instance y, an instance x, and a number of samples $M \in \mathbb{N}$.
Output: An estimate $\widehat{v}$ of $\mathrm{Pr}_{z \sim \mathbf{U}(y)}[\mathcal{L}(z) = \mathcal{L}(x)]$.

1: $\widehat{v} \leftarrow 0$
2: **for** $i = 1$ to M **do**
3:    Sample $z \sim \mathbf{U}(y)$
4:    **if** $\mathcal{L}(z) = \mathcal{L}(x)$ **then**
5:       $\widehat{v} \leftarrow \widehat{v} + 1$
6:    **end if**
7: **end for**
8: **return** $\widehat{v}/M$.

Let S be a random variable corresponding to the set of values k such that Algorithm 1 enters line 13 with $m = k$, and note that if for every k in S it happens that the events

$$A_k := (v_k \geqslant \delta^\star) \text{ and } B_k := (\widetilde{v}_k \geqslant \delta^\star)$$

are equivalent (i.e., either both occur or neither occurs), then the algorithm will succeed, as that would indeed imply that $k^\star$ is the smallest k such that $v_k \geqslant \delta^\star$.

Then, for $k \in [d]$, define events E_k and F_k as follows:

$$E_k := |\delta^\star - v_k| \geqslant \frac{\varepsilon\gamma}{\log d},$$
$$F_k := |\widetilde{v}_k - v_k| \leqslant \frac{\varepsilon\gamma}{\log d}.$$

We claim that if both E_k and F_k hold for some k, then A_k and B_k are equivalent events for that k. Indeed,

$$
\begin{aligned}
A_k &\iff v_k \geqslant \delta^\star \\
&\iff v_k \geqslant \delta^\star + \frac{\varepsilon\gamma}{\log d} && \text{(by } E_k) \\
&\iff v_k - \frac{\varepsilon\gamma}{\log d} \geqslant \delta^\star \\
&\iff \widetilde{v}_k \geqslant \delta^\star && \text{(by } F_k) \\
&\iff B_k.
\end{aligned}
$$

Thus, if we show that E_k and F_k hold with good probability for every $k \in S$, we can conclude the theorem. First, note that because of the condition on the variable sᴛᴇᴘs (lines 10, 11) we have $|S| \leqslant \log d$, allowing us to do a binary search in case the desired events E_k and F_k hold, and preventing the algorithm from looping otherwise; this way the runtime is bounded not only on

10

expectation but deterministically. [6] Then, note that for any k we have

$$\Pr\left[\overline{F_k}\right] \leqslant \Pr\left[\overline{F_k} \mid k \in S\right] = \Pr\left[|\widehat{v}_k(M) - v_k| > \frac{\varepsilon\gamma}{\log d}\right]$$

$$\leqslant \frac{\gamma}{\log d}. \qquad\qquad \text{(by Fact 1)}$$

Because S itself is a random variable, whose size is also a random variable, we need to be careful before applying a union bound or any related tricks. Let us refer to the elements of S as $\{s_1, \ldots, s_\ell\}$, and let us call $F(i)$, for $i \in [\log d]$,[7] to the event F_{s_i} if $i \leqslant \ell$, and to the sample space Ω (i.e., the event that always happens) otherwise. Then, we claim that for any $0 \leqslant i \neq j \leqslant \lceil\log d\rceil$, we have

$$\Pr[F(i) \cap F(j)] = \Pr[F(i)] \cdot \Pr[F(j)], \qquad\qquad (1)$$

as either $\max\{i, j\} \leqslant \ell$, in which case the claim holds by independence (since both events depend only on disjoint sets of independent random samples), or the claim holds trivially since $\Pr[F(i)] = 1$ for $i > \ell$. Therefore, we have

$$\Pr\left[\bigcap_{k\in S} F_k\right] = \Pr\left[F(0) \cap F(1) \cap \cdots \cap F(\lceil\log d\rceil)\right]$$

$$= \prod_{i\in[\log d]} \Pr[F(i)] \qquad\qquad \text{(by Equation (1))}$$

$$\geqslant \left(1 - \frac{\gamma}{\log d}\right)^{\log d} \geqslant 1 - \gamma.$$

We now argue that the event $\bigcap_{k\in S} E_k$ happens with good probability. To see that, note first that for every $k \in [d]$, line 1 implies

$$\Pr[\overline{E_k}] = \Pr\left[\delta^\star \in \left[v_k \pm \frac{\varepsilon\gamma}{\log d}\right]\right] \leqslant \frac{\frac{2\varepsilon\gamma}{\log d}}{2\varepsilon} = \frac{\gamma}{\log d}.$$

Once again, we need to be careful as the events E_k are not independent of S, nor between them this time. Using the law of total probabilities, we have

$$\Pr\left[\bigcap_{k\in S} E_k\right] = \sum_{S'\subseteq[d]} \Pr\left[S = S' \mid \bigcap_{k\in S'} E_k\right] \Pr\left[\bigcap_{k\in S'} E_k\right]$$

$$= \sum_{\substack{S'\subseteq[d] \\ |S'|\leqslant\log d}} \Pr\left[S = S' \mid \bigcap_{k\in S'} E_k\right] \Pr\left[\bigcap_{k\in S'} E_k\right],$$

where we can now effectively use the union bound to say that for any fixed S' with $|S'| \leqslant \log d$ we have

$$\Pr\left[\bigcap_{k\in S'} E_k\right] \geqslant 1 - \gamma.$$

---

[6] For simplicity, we will say $|S| \leqslant \log d$, even though the exact bound for a binary search is $|S| \leqslant \lfloor\log d + 1\rfloor$. This choice naturally has no impact on the asymptotic analysis of the algorithm.

[7] Throughout this proof, we use notation $[\alpha]$, for $\alpha \in \mathbb{R}^{>0}$, to denote the set $\{0, 1, \ldots, \lceil\alpha\rceil - 1, \lceil\alpha\rceil\}$.

Therefore, we conclude that

$$\Pr\left[\bigcap_{k\in S}E_k\right] = \sum_{\substack{S'\subseteq[d]\\|S'|\leq\log d}}\Pr\left[S=S'\mid\bigcap_{k\in S'}E_k\right]\Pr\left[\bigcap_{k\in S'}E_k\right] \geq (1-\gamma)\sum_{\substack{S'\subseteq[d]\\|S'|\leq\log d}}\Pr\left[S=S'\mid\bigcap_{k\in S'}E_k\right]. \quad (2)$$

It is not obvious, however, whether the sum on the right-hand side of Equation (2) equals 1; we will, however, argue that it is at least $1-\gamma$. Recall that $\Pr[F_k|k\notin S]\geq\Pr[F_k|k\in S]$ for any k, from where it follows that for every set S' of size at most $\log d$ we have

$$\Pr\left[\bigcap_{k\in S'}F_k\right]\geq\Pr\left[\bigcap_{k\in S}F_k\right].$$

Then, note that for any index $k\in[d]$, the event F_k is conditionally independent of all events E_j, with $j\in[d]$ given the event $k\in S$. That is,

$$\Pr\left[F_k|k\in S\right] = \Pr\left[F_k|E_j,k\in S\right], \quad\text{for any } j\in[d].$$

We thus deduce that for any fixed S' with $|S'|\leq\log d$ we have

$$\begin{aligned}
\Pr\left[\bigcap_{k\in S'}F_k\mid\bigcap_{k\in S'}E_k\right] &\geq \Pr\left[\bigcap_{k\in S'}F_k\mid\bigcap_{k\in S'}E_k\text{ and }\bigcap_{k\in S'}k\in S\right]\\
&= \Pr\left[\bigcap_{k\in S'}F_k\mid\bigcap_{k\in S'}k\in S\right]\geq\Pr\left[\bigcap_{k\in S}F_k\right]\geq(1-\gamma).
\end{aligned} \quad (3)$$

Then, our key observation is that there is a single value $S^\star\subseteq[d]$, with $|S^\star|\leq\log d$, that the binary search can take if we condition on all the events E_k and F_k happening, since in that case events A_k and B_k coincide. In other words, there exists $S^\star$, with $S^\star\subseteq[d]$ and $|S^\star|\leq\log d$, such that

$$\Pr\left[S=S^\star\mid\bigcap_{k\in S^\star}E_k\cap\bigcap_{k\in S^\star}F_k\right] = 1. \quad (4)$$

We can then argue as follows:

$$\begin{aligned}
\Pr\left[\bigcap_{k\in S}E_k\right] &\geq (1-\gamma)\sum_{\substack{S'\subseteq[d]\\|S'|\leq\log d}}\Pr\left[S=S'\mid\bigcap_{k\in S'}E_k\right] &&\text{(by Equation (2))}\\
&\geq (1-\gamma)\Pr\left[S=S^\star\mid\bigcap_{k\in S^\star}E_k\right]\\
&\geq (1-\gamma)\Pr\left[S=S^\star\text{ and }\bigcap_{k\in S^\star}F_k\mid\bigcap_{k\in S^\star}E_k\right]\\
&= (1-\gamma)\Pr\left[S=S^\star\mid\bigcap_{k\in S^\star}E_k\cap\bigcap_{k\in S^\star}F_k\right]\cdot\Pr\left[\bigcap_{k\in S^\star}F_k\mid\bigcap_{k\in S^\star}E_k\right] &&\text{(Bayes' rule)}\\
&\geq (1-\gamma)^2. &&\text{(by Equations (3) and (4))}
\end{aligned}$$

Therefore, the algorithm will succeed with probability at least

$$\Pr\left[\bigcap_{k \in S} E_k\right] \cdot \Pr\left[\bigcap_{k \in S} F_k\right] \geqslant (1 - \gamma)^3 \geqslant 1 - 3\gamma.$$

The runtime is simply $O(\log d \cdot M \cdot d)$; as (i) the binary search performs $O(\log d)$ steps; (ii) each of the binary search steps requires M samples, and (iii) each sample requires evaluating the model $\mathcal{L}$ and thus takes time $O(d)$. Naturally, running the algorithm with $\gamma' = 1/3 \cdot \gamma$ will yield a success probability of $1 - \gamma$ without changing the asymptotic runtime, and thus we conclude the proof.

□

4 Locally Minimal Probabilistic Explanations

Due to the complexity of finding even subset-minimal δ-SR, Izza et al. (2024) have proposed to study "locally minimal" δ-SR, which are δ-SRs such that the removal of any feature from the explanation would decrease its probabilistic guarantee below δ. Interestingly, we can generalize a proof from Arenas et al. (2022) to show that, over lineal models even in the more general case of product distributions (distributions over $\{0, 1\}^d$ that are products of independent Bernoulli variables of potentially different parameters), every locally minimal δ-SR is a subset-minimal δ-SR. This allows leveraging the previous results of Izza et al. (2024) to subset-minimal δ-SRs in the case of linear models.

Theorem 3. *For linear models, under any product distribution, every locally minimal δ-SR is a subset-minimal δ-SR.*

Proof sketch. Define the *"locality"* gap $\textsc{lgap}(\boldsymbol{y})$ of a locally minimal δ-SR $\boldsymbol{y}$ as the smallest value g such that $|\boldsymbol{y}^\star|_\perp - |\boldsymbol{y}|_\perp = g$ for some $\boldsymbol{y}^\star \subseteq \boldsymbol{y}$ that is a δ-SR. If $g = 0$, then $\boldsymbol{y}$ is globally minimal, and we are done. If g were to be 1, then $\boldsymbol{y}$ would not be locally minimal, a contradiction. Therefore, we can safely assume $g \geqslant 2$ from now on. Let $\mathcal{L}, \boldsymbol{y}$ be such that $\boldsymbol{y}$ is locally minimal δ-SR and $\textsc{lgap}(\boldsymbol{y}) \geqslant 2$. We will find a contradiction by the following method:

- Let $\boldsymbol{y}^\star$ be the δ-SR such that $|\boldsymbol{y}\backslash\boldsymbol{y}^\star| = \textsc{lgap}(\boldsymbol{y})$.

- Every feature in $\boldsymbol{y}\backslash\boldsymbol{y}^\star$ is either "good", if its score is positive, or "bad" if its score is negative.

- Fix any feature i in $\boldsymbol{y}\backslash\boldsymbol{y}^\star$. If i is good, then $\boldsymbol{y}^\star \oplus i$, meaning the partial instance obtained by taking $\boldsymbol{y}$ and setting its i-th feature to x_i, has a probability guarantee greater or equal than that of $\boldsymbol{y}^\star$ (the proof of this fact is very similar to the proof of Lemma 1), and the gap has reduced. On the other hand, if i is bad, then $\boldsymbol{y} \ominus i$, meaning the partial instance obtained from $\boldsymbol{y}$ by setting $y_i = \perp$, has greater-equal probability than $\boldsymbol{y}$, contradicting the fact that $\boldsymbol{y}$ is locally minimal.

□

5 Approximations on the Size of Explanations

A natural question at this point is whether the size of a (δ, ε)-min-SR is necessarily similar to the size of a $(\delta, 0)$-min-SR (i.e., a smallest δ-SR). It turns out that this is not the case, and it can happen that in order to get a slightly better probabilistic guarantee (i.e., $\delta + \varepsilon$ instead of δ), the number of features needed under any explanation significantly increase. In general, if we let $\text{Min}(\mathcal{M}, x, \delta)$ denote the size of the smallest δ-SR for $(\mathcal{M}, x)$, we can prove the following generalization of Example 2.

Proposition 2. *For any $\delta \in (0, 1)$, $\gamma > 0$, and any $\varepsilon > 0$ such that $\delta + \varepsilon \leqslant 1$, there are pairs $(\mathcal{L}, x)$ where $\mathcal{L}$ is a linear model of dimension d, and x an instance of dimension d, such that*

$$\frac{\text{Min}(\mathcal{L}, x, \delta + \varepsilon)}{\text{Min}(\mathcal{L}, x, \delta)} = \Omega\left(d^{\frac{1}{2} - \gamma}\right).$$

As a consequence, we may say informally that approximations on δ do not neccesarily lead to approximations on the explanation size.

6 Conclusions and Future Work

We have proved a positive result for the case of linear models, showing that a (δ, ε)-min-SRs can be computed efficiently, and also a more abstract reason suggesting that linear models might be easier to explain than, e.g., decision trees. However, a variety of natural questions and directions of research remain open. First, in practical terms, even though the runtime of Theorem 2 is polynomial and only has a quasi-linear dependency on d, our future work includes lowering the dependency in $1/\varepsilon$ and $1/\gamma$; on a dataset with $d = 500$, setting $\varepsilon = 0.1$ and $\gamma = 0.01$ is already computationally expensive. We acknowledge, in terms of practical implementations, the work of Bounia and Koriche (2023); Izza et al. (2024) that allows for computing small probabilistic explanations over decision trees significantly faster than the exact SAT approach of Arenas et al. (2022). Similarly, Izza et al. (2023b) showed solid practical results with different kinds of classifiers, including linear models (i.e., Naive Bayes). Despite our results having better theoretical guarantees over linear models, a natural direction of future work is to improve the practical efficiency of our algorithm for high-dimensional models.

Second, our theoretical result has some natural directions for generalization. We considered only binary features, whereas in order to offer a practically useful tool to the community, we will need to understand how to compute (approximate) probabilistic explanations for mixtures real-valued features and categorical features, for example under the "extended linear classifier" definition of Marques-Silva et al. (2020b). Another fascinating theoretical question is handling the generalization of our setting to that of product distributions (i.e., feature i takes value 1 with probability p_i and 0 otherwise) can also be solved efficiently. A straightforward extension of our techniques does not seem to work on such a generalized setting, since the *feature selection* argument of **??** no longer holds. Therefore, we believe that new techniques will be needed.

Third, it would be interesting to allow for a more declarative way of specifying the probabilistic guarantees or constraints on the explanations. While a recent line of research has studied the design of languages for defining explainability queries with a uniform algorithmic treatment (Arenas et al.,

2021; Barceló et al., 2020; Arenas et al., 2024), we are not aware of any work on that line that allows for probabilistic terms.

References

Arenas, M., Báez, D., Barceló, P., Pérez, J., and Subercaseaux, B. (2021). Foundations of Symbolic Languages for Model Interpretability. In *Advances in Neural Information Processing Systems*, volume 34, pages 11690–11701. Curran Associates, Inc.

Arenas, M., Barceló, P., Romero Orth, M., and Subercaseaux, B. (2022). On computing probabilistic explanations for decision trees. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28695–28707. Curran Associates, Inc.

Arenas, M., Barceló, P., Bustamante, D., Caraball, J., and Subercaseaux, B. (2024). A Uniform Language to Explain Decision Trees. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 60–70.

Barceló, P., Monet, M., Pérez, J., and Subercaseaux, B. (2020). Model interpretability through the lens of computational complexity. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15487–15498. Curran Associates, Inc.

Barceló, P., Pérez, J., and Subercaseaux, B. (2020). Foundations of languages for interpretability and bias detection. *AFCI workshop at NeurIPS 2020. Algorithmic Fairness through the Lens of Causality and Interpretability*.

Blanc, G., Lange, J., and Tan, L.-Y. (2021). Provably efficient, succinct, and precise explanations. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Bounia, L. and Koriche, F. (2023). Approximating probabilistic explanations via supermodular minimization. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org.

Darwiche, A. and Hirth, A. (2020). On the Reasons Behind Decisions. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, pages 712–720.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30.

Ignatiev, A., Narodytska, N., Asher, N., and Marques-Silva, J. (2021). From contrastive to abductive explanations and back again. In Baldoni, M. and Bandini, S., editors, *AIxIA 2020 – Advances in Artificial Intelligence*, Lecture Notes in Computer Science, page 335–355, Cham. Springer International Publishing.

Izza, Y., Huang, X., Ignatiev, A., Narodytska, N., Cooper, M., and Marques-Silva, J. (2023a). On computing probabilistic abductive explanations. *International Journal of Approximate Reasoning*, 159:108939.

Izza, Y., Huang, X., Ignatiev, A., Narodytska, N., Cooper, M., and Marques-Silva, J. (2023b). On computing probabilistic abductive explanations. *International Journal of Approximate Reasoning*, 159:108939.

Izza, Y., Ignatiev, A., and Marques-Silva, J. (2020). On Explaining Decision Trees.

Izza, Y., Ignatiev, A., Narodytska, N., Cooper, M. C., and Marques-Silva, J. (2021). Efficient explanations with relevant sets. *ArXiv*, abs/2106.00546.

Izza, Y., Meel, K. S., and Marques-Silva, J. (2024). Locally-minimal probabilistic explanations.

Komjáth, P. (2013). Upper limit on the central binomial coefficient. MathOverflow. URL:https://mathoverflow.net/q/133742 (version: 2013-07-01).

Kozachinskiy, A. (2023). Inapproximability of sufficient reasons for decision trees. arXiv:2304.02781 [cs].

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., and Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. arXiv:1902.00006 [cs, stat].

Marques-Silva, J., Gerspacher, T., Cooper, M. C., Ignatiev, A., and Narodytska, N. (2020a). Explaining naive bayes and other linear classifiers with polynomial time and delay. *CoRR*, abs/2008.05803.

Marques-Silva, J., Gerspacher, T., Cooper, M. C., Ignatiev, A., and Narodytska, N. (2020b). Explaining naive bayes and other linear classifiers with polynomial time and delay. In *NeurIPS*.

Marques-Silva, J. and Ignatiev, A. (2022). Delivering trustworthy ai through formal xai. In *AAAI*.

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

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv:1802.00682 [cs].

Subercaseaux, B. (2020). Model interpretability through the lens of computational complexity. Master's thesis, Universidad de Chile.

Wäldchen, S., MacDonald, J., Hauch, S., and Kutyniok, G. (2021). The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res.*, 70:351–387.

Appendix

.1 Calculation for Example 2

Consider the following version of Chernoff bound.

Lemma 2 (Chernoff bound). *Let X be a finite sum of independent Bernoulli variables, with $\mathbb{E}[X] = \mu > 0$. Then for any $t \geqslant 0$ we have*

$$\Pr\left[\,|X - \mu| \geqslant t\right] \leqslant 2\exp\left(\frac{-t^2}{3\mu}\right).$$

If we define the Bernoulli variables $X_i := (z_i = 1)$ for $z \sim \mathbf{U}(y)$ (the uniform distribution over $\mathrm{Comp}(y)$), then the variables $X_2, \dots, X_{1000}$ are identical independent Bernoulli variables, and if $X = \sum_{i=2}^{1000} X_i$ we have

$$\mu := \mathbb{E}[X] = 999\mathbb{E}[X_2] = \frac{999}{2},$$

as each X_i has expectation $\frac{1}{2}$ becuase $\mathbf{U}$ is the uniform distribution. Then, using Lemma 2, we have that

$$\Pr\left[X < 250\right] \leqslant 2\exp\left(\frac{-(445 - 250)^2 \cdot 2}{3 \cdot 999}\right)$$

$$\approx 2\exp(-32.29...)$$

$$< 2 \cdot 10^{14}.$$

To conclude, note that

$$\Pr_{z \sim \mathbf{U}(y)}\left[\mathcal{L}(z) = 1\right] = \Pr_{z \sim \mathbf{U}(y)}\left[z \cdot w \geqslant 1250\right]$$

$$= \Pr_{z \sim \mathbf{U}(y)}\left[1000 + \sum_{i=2}^{1000} z_i \geqslant 1250\right]$$

$$= 1 - \Pr\left[\sum_{i=2}^{1000} z_i < 250\right]$$

$$= 1 - \Pr\left[X < 250\right] > 1 - 2 \cdot 10^{14} > 0.999999.$$

.2 Proposition 2

In order to prove Proposition 2 we will need some auxiliary lemmas concerning binomial distributions. The only previous result we will need is a simple upper bound on the central binomial coefficient, attributed to Erdős (Komjáth, 2013).

Lemma 3 (Erdős' bound (see (Komjáth, 2013))). *For every $n \geqslant 1$ we have*

$$\binom{2n}{n} \leqslant \frac{4^n}{\sqrt{2n + 1}}.$$

We can now proceed to prove our auxiliary lemmas.

Lemma 4. *Let $P(n,k)$ be the probability that a binomial random variable $X \sim Bin(n,1/2)$ is at least k, and define*

$$g(n) = \max_{k \in \{1,\dots,n\}} P(n-1,k-1) - P(n,k).$$

Then, $0 \leqslant g(n) \leqslant 1/\sqrt{n}$ for every $n \geqslant 2$.

Proof. We can think of $P(n,k)$ as the probability of getting at least k heads out of n tosses of a fair coin. By casing on the value (heads or tails) of the first coin we can see that for $n \geqslant k \geqslant 1$ we have

$$P(n,k) = \frac{1}{2}P(n-1,k-1) + \frac{1}{2}P(n-1,k),$$

which we can rearrange into

$$P(n,k) = P(n-1,k-1) + \frac{1}{2}\big[P(n-1,k) - P(n-1,k-1)\big],$$

and thus

$$P(n-1,k-1) - P(n,k) = \frac{1}{2}\big[P(n-1,k-1) - P(n-1,k)\big].$$

But $P(n-1,k-1) - P(n-1,k)$ is the probability of getting exactly $k-1$ heads in $n-1$ tosses, which is $\frac{1}{2^{n-1}}\binom{n-1}{k-1}$, and clearly positive. Hence, using [Lemma 3](), we have

$$0 \leqslant P(n-1,k-1) - P(n,k) \leqslant \frac{1}{2^n}\binom{n-1}{k-1} \leqslant \frac{1}{2^n}\binom{n-1}{\lfloor(n-1)/2\rfloor} \leqslant \frac{1}{2^n}\frac{2^n}{\sqrt{2n-1}} \leqslant \frac{1}{\sqrt{n}}.$$

$\square$

Lemma 5 (Binomial Approximation). *Let $P(n,k)$ be the probability that a binomial random variable $X \sim Bin(n,1/2)$ is at least k. Then, given a probability $\delta \in (0,1)$ and a targer error $\varepsilon > 0$, there exist values $n \leqslant \varepsilon^{-2}$ and $k \leqslant n$ such that*

$$|P(n,k) - \delta| < \varepsilon.$$

Proof. Fix any value of n and note that $P(n,k-1) - P(n,k)$ is the probability of getting exactly $k-1$ heads in n tosses of a fair coin. Thus, for any $k \in \{1,\dots,n\}$ we can use [Lemma 3]() to say

$$P(n,k-1) - P(n,k) = \frac{1}{2^n}\binom{n}{k-1} \leqslant \frac{1}{2^n}\binom{n}{\lfloor n/2\rfloor} \leqslant \frac{1}{2^n}\frac{2^n}{\sqrt{2n+1}} \leqslant \frac{1}{\sqrt{n}}. \tag{5}$$

Now, choose a large enough n so that $\frac{1}{\sqrt{n}} < \varepsilon$. Then, choose k as the largest integer such that $P(n,k) \geqslant \delta$, which exists since $P(n,0) = 1$ and $P(n,n+1) = 0$. This definition of k yields

$$P(n,k) \geqslant \delta > P(n,k+1),$$

and using [Equation (5)]() we can see that

$$|P(n,k+1) - \delta| = \delta - P(n,k+1) \leqslant P(n,k) - P(n,k+1) \leqslant \frac{1}{\sqrt{n}} < \varepsilon,$$

which completes the proof.

$\square$

We are finally ready to prove Proposition 2.

Proof of Proposition 2. Let δ and ε be the values in the statement of the proposition. Let n and m be large integers whose precise value will be determined later. We will construct a linear model $\mathcal{L}$ of dimension $d = n + 1$, with threshold $t = 2$, and weights $w_1 = 1$, $w_i = 1/m$ for $i \in \{2, \ldots, d\}$. Let $x = (1, 1, \ldots, 1) \in \{0, 1\}^d$ be the instance to explain, and $y = (1, \perp, \perp, \ldots, \perp) \subseteq x$ be a partial instance. Now, use Lemma 5 with parameters $\delta' = \delta + \varepsilon/4$ and $\varepsilon' = \varepsilon/4$ to choose values for n, m such that

$$\left| P(n, m) - \delta' \right| < \varepsilon',$$

with $m \leqslant n$. Note now that $\mathcal{L}(x) = 1$, and that

$$\Pr_{z \sim \mathbf{U}(y)} [\mathcal{L}(z) = 1] = P(n, m), \tag{6}$$

since in order for a completion of y to have $\mathcal{L}(z) = 1$ it must have at least m features (besides the 1st one) set to 1, as that way its value according to the weights w_i will be at least $1 + m \cdot 1/m = 2 = t$. As we have

$$\left| P(n, m) - (\delta + \varepsilon/4) \right| < \varepsilon/4,$$

we can deduce

$$\delta \leqslant P(n, m) \leqslant \delta + \varepsilon/2.$$

Combining our last equation with Equation (6) we have that y is a δ-SR for x of size 1. Now, let us see how many more features need to be added to y in order to get a $(\delta + \varepsilon)$-SR. Suppose that we add k features to y (as all features are equal, we add any k of them) obtaining a partial instance y'. Then, we have that

$$\Pr_{z \sim \mathbf{U}(y')} [\mathcal{L}(z) = 1] = P(n - k, m - k),$$

as now k features have been set to 1. We can now use Lemma 4 k times in a row:

$$
\begin{aligned}
P(n - k, m - k) &\leqslant P(n - k + 1, m - k + 1) + \frac{1}{\sqrt{n - k}} \\
&\leqslant P(n - k + 2, m - k + 2) + \frac{1}{\sqrt{n - k}} + \frac{1}{\sqrt{n - k + 1}} \\
&\leqslant P(n - k + 2, m - k + 2) + \frac{2}{\sqrt{n - k}} \\
&\ \ \vdots \\
&\leqslant P(n, m) + \frac{k}{\sqrt{n - k}}.
\end{aligned}
$$

Therefore, we have that

$$P(n - k, m - k) \leqslant P(n, m) + \frac{k}{\sqrt{n - k}} \leqslant \delta + \varepsilon/2 + \frac{k}{\sqrt{n - k}}.$$

It now remains to show that, unless k is large enough (relatively to n), the right-hand side of the previous inequality is less than $\delta + \varepsilon$. Indeed, take $k = n^{1/2-\gamma}$ for any positive γ. Then, we want to show that for sufficiently large n we have

$$\frac{k}{\sqrt{n-k}} < \varepsilon/2.$$

To do so, note that

$$\frac{k}{\sqrt{n-k}} < \frac{k}{\sqrt{n}} = \frac{n^{1/2-\gamma}}{n^{1/2}} = n^{-\gamma},$$

from where taking $n > \left(\frac{2}{\varepsilon}\right)^{1/\gamma}$ suffices to complete the proof. $\qquad\square$

.3 Proposition 1

Proof of Proposition 1. The proof is a twist on Barceló et al. (2020, Lemma 28); let $(s_1, \ldots, s_n, T) \in \mathbb{N}^{n+1}$ be an instance of the #P-complete problem #Knapsack, that consists on counting the number of sets $S \subseteq \{s_1, \ldots, s_n\}$ such that $\sum_{s \in S} s \leqslant T$. We can assume that $\sum_{i=1}^{n} s_i > T$, as otherwise the #Knapsack instance is trivial. Then, let $\mathcal{L}$ be a linear model with weights $w_i = s_i$, and threshold $t = T + 1$. Now, consider the problem of deciding whether #Knapsack$(s_1, \ldots, s_n, T) \geqslant m$ for an input m, which cannot be solved in polynomial time unless P = #P. Let $x = (1, 1, \ldots, 1)$, and $\delta = \frac{m}{2^n}$. We claim that $(\mathcal{L}, x, \delta, k = 0)$ is a Yes-instance for Uniform-Min-δ-SR(Linear) if and only if #Knapsack$(s_1, \ldots, s_n, T) \geqslant m$. First, note that $\mathcal{L}(x) = 1$, since $\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} s_i \geqslant T + 1 = t$. For every set $S \subseteq \{s_1, \ldots, s_n\}$ such that $\sum_{s \in S} s \leqslant T$, its complement $\overline{S} := \{s_1, \ldots, s_n\} \backslash S$ holds $\sum_{s \in \overline{S}} s > T$, and as all values are integers, this implies as well

$$\sum_{s \in \overline{S}} s \geqslant T + 1 = t.$$

To each such set $\overline{S}$, we associate the instance $z(\overline{S})$ defined as

$$z(\overline{S})_i = \begin{cases} 1 & \text{if } s_i \in \overline{S} \\ 0 & \text{otherwise.} \end{cases}$$

Now note that

$$\mathcal{L}(z(\overline{S})) = \begin{cases} 1 & \text{if } \sum_{s_i \in \overline{S}} w_i \geqslant T + 1 \\ 0 & \text{otherwise} \end{cases} = 1,$$

and thus there is a bijection between the sets S whose sum is at most T and the instances $z(\overline{S})$ such that $\mathcal{L}(z(\overline{S})) = 1 = \mathcal{L}(x)$. To conclude, simply note that the previous bijection implies

$$\Pr_{z \sim U(\perp^d)} \left[\mathcal{L}(z) = 1 \right] = \frac{\#\text{Knapsack}(s_1, \ldots, s_n, T)}{2^n},$$

and thus the "empty explanation" $\perp^d := (\perp, \perp, \ldots, \perp)$ has probability at least δ if and only if

$$\#\text{Knapsack}(s_1, \ldots, s_n, T) \geqslant m.$$

As the empty explanation is the only one with size $\leqslant 0$, we conclude the proof.

$\qquad\square$

.4 Lemma 1 and Theorem 4

Let us state the result about locally minimal δ-SRs as a theorem, so we can proceed to prove it. In order to state it, however, we need to define what we mean by an arbitrary product distribution. Consider Bernoulli variables $X_1, \ldots, X_d$ with probabilities $p_1, \ldots, p_d$ respectively, and let us denote by $\mathbf{D} = X_1 \times \cdots \times X_d$ their joint distribution, which we will call a product distribution. Then, $\mathbf{D}$ induces a distribution over the set $\{0, 1\}^d$ by the rule

$$\Pr_{z \sim \mathbf{D}}[z = x] = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}, \quad \forall x \in \{0, 1\}^d.$$

Naturally, this induces a distribution over $\mathrm{Comp}(y)$ for any partial instance y as follows:

$$\Pr_{z \sim \mathbf{D}(y)}[z = x] = \frac{\Pr_{z \sim \mathbf{D}}[z = x]}{\sum_{w \in \mathrm{Comp}(y)} \Pr_{z \sim \mathbf{D}}[z = x]} \quad \forall x \in \mathrm{Comp}(y),$$

and naturally $\Pr_{z \sim \mathbf{D}(y)}[z = x] = 0$ if $x \notin \mathrm{Comp}(y)$. With this notation, we can make a precise theorem statement.

Theorem 4. *For linear models, under any product distribution, every locally minimal δ-SR is a subset-minimal δ-SR.*

Furthermore, we will prove a stronger lemma than Lemma 1 stated in the main body.

Lemma 1. *Given a linear model $\mathcal{L}$, and an instance x, if $y^{(0)}, \ldots, y^{(d)}$ are the partial instances of x such that $y^{(k)} \subseteq x$ is defined only in the top k features of maximum score, then*

$$\Pr_{z \sim \mathbf{U}(y^{(k+1)})}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \Pr_{z \sim \mathbf{U}(y^{(k)})}[\mathcal{L}(z) = \mathcal{L}(x)]$$

for all $k \in \{1, \ldots, d-1\}$, and naturally,

$$\Pr_{z \sim \mathbf{U}(y^{(d)})}[\mathcal{L}(z) = \mathcal{L}(x)] = 1.$$

Moreover, $\mathrm{MIN}(\mathcal{L}, x, \delta) = k$ if and only if $y^{(k)}$ is a δ-SR for x but $y^{(k-1)}$ is not.

Before we prove Lemma 1, let us prove an auxiliary lemma that will also be useful for proving Theorem 3. Given an instance x to explain, and a partial instance $y \subseteq x$, such that $y_i = \bot$, we define the partial instance $y \oplus i$ by:

$$(y \oplus i)_j = \begin{cases} y_j & \text{if } j \neq i \\ x_i & \text{otherwise.} \end{cases}$$

Lemma 6. *Let $\mathcal{L}$ be a linear model, x an instance and $y \subseteq x$ a partial instance. Assume a product distribution $\mathbf{D}$. Then, if i is a feature such that $y_i = \bot$, and the feature score s_i holds $s_i \geqslant 0$, we have*

$$\Pr_{z \in \mathbf{D}(y \oplus i)}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \Pr_{z \in \mathbf{D}(y)}[\mathcal{L}(z) = \mathcal{L}(x)].$$

Proof of Lemma 6. Let $(w_1, \ldots, w_d)$ be the weights of $\mathcal{L}$ and t its threshold. Then, let us assume without loss of generality that $\mathcal{L}(x) = 1$, as the case $\mathcal{L}(x) = 0$ is analogous. We thus have that

$$s_i = w_i \cdot (2x_i - 1) \geqslant 0,$$

from where $s_i = w_i$ if $x_i = 1$ and $s_i = -w_i$ if $x_i = 0$. Let us denote by S the set of features j such that $y_j = x_j \neq \bot$, and define

$$t' := t - \sum_{j \in S} y_j w_j.$$

We can then rewrite the probability of interest as

$$\Pr_{z \sim \mathbf{D}(y)}[\mathcal{L}(z) = \mathcal{L}(x)] = \Pr_{z \sim \mathbf{D}(y)}[\mathcal{L}(z) = 1] = \Pr_{z \sim \mathbf{D}(y)}\left[\sum_{j \notin S} z_j w_j \geqslant t'\right].$$

Let us define the following two amounts:

$$\mathcal{A} := \Pr_{z \sim \mathbf{D}(y)}\left[\sum_{j \notin S, j \neq i} z_j w_j \geqslant t' - w_i\right],$$

$$\mathcal{B} := \Pr_{z \sim \mathbf{D}(y)}\left[\sum_{j \notin S, j \neq i} z_j w_j \geqslant t'\right].$$

If p_i is the probability of feature i under $\mathbf{D}$, then we have

$$\Pr_{z \sim \mathbf{D}(y)}[\mathcal{L}(z) = \mathcal{L}(x)] = \Pr_{z \sim \mathbf{D}(y)}\left[\sum_{j \notin S} z_j w_j \geqslant t'\right] = p_i \cdot \mathcal{A} + (1 - p_i) \cdot \mathcal{B}.$$

Now we proceed by cases on x_i. If $x_i = 1$, then $s_i = w_i$, and thus we know $w_i \geqslant 0$, from where $t' - w_i \leqslant t'$ and thus $\mathcal{A} \geqslant \mathcal{B}$. Moreover, as $x_i = 1$, we conclude

$$\Pr_{z \sim \mathbf{D}(y \oplus i)}[\mathcal{L}(z) = \mathcal{L}(x)] = \mathcal{A}$$

$$= p_i \cdot \mathcal{A} + (1 - p_i) \cdot \mathcal{A}$$
$$\geqslant p_i \cdot \mathcal{A} + (1 - p_i) \cdot \mathcal{B}$$
$$= \Pr_{z \sim \mathbf{D}(y)}[\mathcal{L}(z) = \mathcal{L}(x)].$$

Similarly, if $x_i = 0$, then $s_i = -w_i$, and thus we know $w_i \leqslant 0$, from where $t' - w_i \geqslant t'$ and thus $\mathcal{A} \leqslant \mathcal{B}$. As $x_i = 0$, we conclude

$$\Pr_{z \sim \mathbf{D}(y \oplus i)}[\mathcal{L}(z) = \mathcal{L}(x)] = \mathcal{B}$$

$$= p_i \cdot \mathcal{B} + (1 - p_i) \cdot \mathcal{B}$$
$$\geqslant p_i \cdot \mathcal{A} + (1 - p_i) \cdot \mathcal{B}$$
$$= \Pr_{z \sim \mathbf{D}(y)}[\mathcal{L}(z) = \mathcal{L}(x)].$$

This concludes the proof. $\qquad \square$

Similarly, for any partial instance $\boldsymbol{y}$ such that $y_i \neq \perp$, we can define the partial instance $\boldsymbol{y} \ominus i$ as

$$(\boldsymbol{y} \ominus i)_j = \begin{cases} y_j & \text{if } j \neq i \\ \perp & \text{otherwise.} \end{cases}$$

The proof of Lemma 6, but reversing signs, yields the following lemma.

Lemma 7. *Let $\mathcal{L}$ be a linear model, $\boldsymbol{x}$ an instance and $\boldsymbol{y} \subseteq \boldsymbol{x}$ a partial instance. Assume a product distribution $\mathbf{D}$. Then, if i is a feature such that $y_i \neq \perp$, and the feature score s_i holds $s_i \leqslant 0$, we have*

$$\Pr_{\boldsymbol{z} \in \mathbf{D}(\boldsymbol{y} \ominus i)}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] \geqslant \Pr_{\boldsymbol{z} \in \mathbf{D}(\boldsymbol{y})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})].$$

Now, in order to prove Lemma 1, which makes two claims, we will split it into two separate lemmas.

Lemma 8 (Part 1 of Lemma 1)**.** *Given a linear model $\mathcal{L}$, and an instance $\boldsymbol{x}$, if $\boldsymbol{y}^{(0)}, \ldots, \boldsymbol{y}^{(d)}$ are the partial instances of $\boldsymbol{x}$ such that $\boldsymbol{y}^{(k)} \subseteq \boldsymbol{x}$ is defined only in the top k features of maximum score, then*

$$\Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k+1)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] \geqslant \Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})]$$

for all $k \in \{0, \ldots, d-1\}$, and naturally,

$$\Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(d)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] = 1.$$

Proof of Lemma 8. Let us assume without loss of generality that the features are already sorted decreasingly in terms of score, so

$$s_1 \geqslant s_2 \geqslant \cdots \geqslant s_d.$$

This way, we have that $\boldsymbol{y}^{(k)} \subseteq \boldsymbol{x}$ is defined as follows:

$$y_i^{(k)} = \begin{cases} x_i & \text{if } i \leqslant k \\ \perp & \text{otherwise.} \end{cases}$$

The proof now requires considering two cases. First, if $s_{k+1} \geqslant 0$, then we can apply Lemma 6 to conclude that

$$\Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k+1)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] \geqslant \Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})].$$

We will now show that if $s_{k+1} < 0$, then

$$\Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k+1)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] = 1,$$

which will be enough to conclude. Indeed, as $s_{k+1} < 0$, we have that

$$0 > s_{k+2} \geqslant s_{k+3} \geqslant \cdots \geqslant s_d,$$

from where we can repeatedly apply Lemma 7 to deduce

$$\Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k+1)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] \geqslant \Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(k+2)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] \geqslant \cdots \geqslant \Pr_{\boldsymbol{z} \sim \mathbf{U}(\boldsymbol{y}^{(d)})}[\mathcal{L}(\boldsymbol{z}) = \mathcal{L}(\boldsymbol{x})] = 1.$$

This concludes the proof. □

Lemma 9 (Part 2 of Lemma 1). *Given a linear model $\mathcal{L}$, and an instance x, if $y^{(0)}, \ldots, y^{(d)}$ are the partial instances of x such that $y^{(k)} \subseteq x$ is defined only in the top k features of maximum score, then $\text{Min}(\mathcal{L}, x, \delta) = k$ if and only if $y^{(k)}$ is a δ-SR for x but $y^{(k-1)}$ is not.*

In order to prove Lemma 9, we will use a separate lemma. Let us define, for every $k \in [d]$ the set P_k as the set of partial instances $y \subseteq x$ such that y has k defined features.

Lemma 10. *For any $k \in [d]$, we have*

$$\Pr_{z \sim \mathbf{U}(y^{(k)})}[\mathcal{L}(z) = \mathcal{L}(x)] = \max_{y \in P_k} \Pr_{z \sim \mathbf{U}(y)}[\mathcal{L}(z) = \mathcal{L}(x)].$$

Let us show immediately how Lemma 9 can be proved using Lemma 10.

Proof of Lemma 9. For the forward direction, assume that $\text{Min}(\mathcal{L}, x, \delta) = k$. Then, by definition, we have that there exists a δ-SR $y^\star$ for x such that $y^\star$ has k defined features. By Lemma 10, we have that

$$\Pr_{z \sim \mathbf{U}(y^{(k)})}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \Pr_{z \sim \mathbf{U}(y^\star)}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \delta,$$

and thus $y^{(k)}$ is a δ-SR for x. On the other hand, if $y^{(k-1)}$ were to be a δ-SR for x, then we would have $\text{Min}(\mathcal{L}, x, \delta) \leqslant k-1$, a contradiction. For the backward direction, assume that $y^{(k)}$ is a δ-SR for x but $y^{(k-1)}$ is not. Then, by Lemma 10, we have that

$$\delta > \Pr_{z \sim \mathbf{U}(y^{(k-1)})}[\mathcal{L}(z) = \mathcal{L}(x)] = \max_{y \in P_{k-1}} \Pr_{z \sim \mathbf{U}(y)}[\mathcal{L}(z) = \mathcal{L}(x)],$$

from where $\text{Min}(\mathcal{L}, x, \delta) > k-1$, and because $y^{(k)}$ is a δ-SR for x, we have $\text{Min}(\mathcal{L}, x, \delta) \leqslant k$; we conclude that $\text{Min}(\mathcal{L}, x, \delta) = k$. $\square$

It thus only remains to prove Lemma 10.

Proof of Lemma 10. Let $w_1, \ldots, w_d$ be the weights of $\mathcal{L}$, and t its threshold. Let us use the $\oplus, \ominus$ notation defined in Lemmas 6 and 7. We will prove something slightly stronger than Lemma 10: that if i and j are features such that $s_i \leqslant s_j$, then for any partial instance y such that $y_i \neq \perp$ and $y_j = \perp$, we have

$$\Pr_{z \sim \mathbf{U}(y \ominus i \oplus j)}[\mathcal{L}(z) = \mathcal{L}(x)] \geqslant \Pr_{z \sim \mathbf{U}(y)}[\mathcal{L}(z) = \mathcal{L}(x)].$$

If we prove this, then we can apply it repeatedly to deduce Lemma 10. To prove the claim, we start by defining

$$S = \{\ell \mid y_\ell \neq \perp\} \setminus \{i\},$$

and

$$t' = t - \sum_{\ell \in S} y_\ell w_\ell.$$

We will also assume without loss of generality that $\mathcal{L}(x) = 1$ since the other case is analogous. We can then rewrite the probabilities of interest as follows, using notation $\bar{S} := [d] \setminus S$:

$$\Pr_{z \sim \mathbf{U}(y \ominus i \oplus j)}[\mathcal{L}(z) = \mathcal{L}(x)] = \Pr_{z \sim \mathbf{U}(y \ominus i \oplus j)}\left[\sum_{\ell \in \bar{S} \setminus \{i,j\}} z_\ell w_\ell + x_j w_j + z_i w_i \geqslant t'\right],$$

$$\Pr_{z \sim \mathbf{U}(y)}[\mathcal{L}(z) = \mathcal{L}(x)] = \Pr_{z \sim \mathbf{U}(y)}\left[\sum_{\ell \in \bar{S} \setminus \{i,j\}} z_\ell w_\ell + x_i w_i + z_j w_j \geqslant t'\right].$$

Let us write $t^\star = t' - \sum_{\ell \in \bar{S} \setminus \{i,j\}} z_\ell w_\ell$, and note that $t^\star$ is a random variable. With this notation, it remains to prove that

$$\Pr_{z_i, t^\star}[z_i w_i + x_j w_j \geqslant t^\star] \geqslant \Pr_{z_i, t^\star}[z_j w_j + x_i w_i \geqslant t^\star]$$

$$\Longleftrightarrow \frac{1}{2} \Pr_{t^\star}[w_i + x_j w_j \geqslant t^\star] + \frac{1}{2} \Pr_{t^\star}[x_j w_j \geqslant t^\star] \geqslant \frac{1}{2} \Pr_{t^\star}[w_j + x_i w_i \geqslant t^\star] + \frac{1}{2} \Pr_{t^\star}[x_i w_i \geqslant t^\star]$$

$$\Longleftrightarrow \Pr_{t^\star}[w_i + x_j w_j \geqslant t^\star] + \Pr_{t^\star}[x_j w_j \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j + x_i w_i \geqslant t^\star] + \Pr_{t^\star}[x_i w_i \geqslant t^\star].$$

We will prove the last inequality by cases, recalling that $s_j \geqslant s_i$ and thus $w_j(2x_j - 1) \geqslant w_i(2x_i - 1)$.

- (**Case 1:** $x_i = 1, x_j = 1$) The desired inequality is

$$\Pr_{t^\star}[w_i + w_j \geqslant t^\star] + \Pr_{t^\star}[w_j \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j + w_i \geqslant t^\star] + \Pr_{t^\star}[w_i \geqslant t^\star]$$

$$\Longleftrightarrow \Pr_{t^\star}[w_j \geqslant t^\star] \geqslant \Pr_{t^\star}[w_i \geqslant t^\star],$$

  which is true since $s_j \geqslant s_i$ implies $w_j \geqslant w_i$ given $x_i = x_j = 1$.

- (**Case 2:** $x_i = 1, x_j = 0$) The desired inequality is

$$\Pr_{t^\star}[w_i \geqslant t^\star] + \Pr_{t^\star}[0 \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j + w_i \geqslant t^\star] + \Pr_{t^\star}[w_i \geqslant t^\star]$$

$$\Longleftrightarrow \Pr_{t^\star}[0 \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j + w_i \geqslant t^\star],$$

  which is true since $s_j \geqslant s_i$ implies $-w_j \geqslant w_i$ given $x_i = 1, x_j = 0$, and thus $w_i + w_j \leqslant 0$.

- (**Case 3:** $x_i = 0, x_j = 1$) The desired inequality is

$$\Pr_{t^\star}[w_i + w_j \geqslant t^\star] + \Pr_{t^\star}[w_j \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j \geqslant t^\star] + \Pr_{t^\star}[0 \geqslant t^\star],$$

$$\Longleftrightarrow \Pr_{t^\star}[w_i + w_j \geqslant t^\star] \geqslant \Pr_{t^\star}[0 \geqslant t^\star],$$

  which is true since $s_j \geqslant s_i$ implies $w_j \geqslant -w_i$ given $x_i = 0, x_j = 1$, and thus $w_i + w_j \geqslant 0$.

- (**Case 4:** $x_i = 0, x_j = 0$) The desired inequality is

$$\Pr_{t^\star}[w_i \geqslant t^\star] + \Pr_{t^\star}[0 \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j \geqslant t^\star] + \Pr_{t^\star}[0 \geqslant t^\star],$$

$$\Longleftrightarrow \Pr_{t^\star}[w_i \geqslant t^\star] \geqslant \Pr_{t^\star}[w_j \geqslant t^\star],$$

  which is true since $s_j \geqslant s_i$ implies $-w_j \geqslant -w_i$ given $x_i = x_j = 0$, and thus $w_i \geqslant w_j$.

$\square$

Lemma 1 now follows directly from Lemma 8 and Lemma 9, and the sketch proof of Theorem 3 can be completed as we now have proved Lemma 6 and Lemma 7.