Explaining k-Nearest Neighbors: Abductive and Counterfactual Explanations

Pablo Barceló
IMC PUC, IMFD, CENIA
pbarcelo@ing.uc.cl

Alexander Kozachinskiy
CENIA
alexander.kozachinskyi@cenia.cl

Miguel Romero Orth
DCC PUC, CENIA
mgromero@uc.cl

Bernardo Subercaseaux
Carnegie Mellon University
bersub@cmu.edu

José Verschae
IMC PUC
jverschae@uc.cl

December 12, 2024

Abstract

Despite the wide use of k-Nearest Neighbors as classification models, their explainability properties remain poorly understood from a theoretical perspective. While nearest neighbors classifiers offer interpretability from a *"data perspective"*, in which the classification of an input vector $\bar{x}$ is explained by identifying the vectors $\bar{v}_1, \dots, \bar{v}_k$ in the training set that determine the classification of $\bar{x}$, we argue that such explanations can be impractical in high-dimensional applications, where each vector has hundreds or thousands of features and it is not clear what their relative importance is. Hence, we focus on understanding nearest neighbor classifications through a *"feature perspective"*, in which the goal is to identify how the values of the features in $\bar{x}$ affect its classification. Concretely, we study *abductive explanations* such as "minimum sufficient reasons", which correspond to sets of features in $\bar{x}$ that are enough to guarantee its classification, and *counterfactual explanations* based on the minimum distance feature changes one would have to perform in $\bar{x}$ to change its classification. We present a detailed landscape of positive and negative complexity results for counterfactual and abductive explanations, distinguishing between discrete and continuous feature spaces, and considering the impact of the choice of distance function involved. Finally, we show that despite some negative complexity results, Integer Quadratic Programming and SAT solving allow for computing explanations in practice.

1 Introduction

Nearest Neighbor classification. k-Nearest Neighbor (k-NN) classification is one of the most widely used supervised learning techniques [14]. In k-NN classification, we assume a set of points S over a metric space, where each point has already been labeled as either *positive* or *negative*. Then, a new point $\bar{x}$ is classified as either positive or negative by taking the majority label of its k closest neighbors in S. The study of k-NN classification has been a recurring focus in the data management community, encompassing extensive research on its behavior in high-dimensional spaces [13, 30, 56] and its properties when dealing with uncertain data [1, 2, 22]. Considerable effort has also been directed toward the development of efficient algorithms and data structures to enable scalable NN queries [3, 53]. As of late, k-NN has also become key to several search and retrieval problems in vector databases [40]. For example, in *Retrieval-Augmented Generation* (RAG) systems, the goal is to identify the most relevant sections of a document for a given query. This is achieved by performing a nearest-neighbor query within a textual-embedding space.

Formal explainability. Emerging data-driven applications, particularly those leveraging machine learning systems, are introducing new demands on classification methods. One of the most critical requirements is *explainability*: in many high-stakes applications, it is not enough for classifiers to be accurate; they must also provide clear and understandable explanations for their decisions [7]. A significant milestone in this field has been the development of *formal* frameworks for explainability. The advantages of adopting such a principled approach have been comprehensively outlined in a recent survey [48]. Two prominent examples of this methodology are:

- *Abductive explanations:* These aim to identify a small subset of components in the input $\bar{x}$ that is sufficient to justify the classifier's output for $\bar{x}$ [18,35,51]. More formally, an abductive explanation for $\bar{x}$ with respect to a given classifier is a subset X of the components of $\bar{x}$ such that every input $\bar{y}$ that coincides with $\bar{x}$ over the components in X is classified in the same way by the classifier. Abductive explanations are also called *sufficient reasons* [55]. One then aims to find sufficient reasons for $\bar{x}$ that are *minimum* in terms of their cardinality.

- *Contrastive explanations:* These focus on the robustness of a classification, examining how much a given point $\bar{x}$ must be altered to change the output of the classifier [8,25]. More formally, a *contrastive explanation at distance p from $\bar{x}$*, with respect to a given classifier, is another input $\bar{y}$ such that $\|\bar{x} - \bar{y}\| \leq p$ and $\bar{y}$ is classified differently from $\bar{x}$.

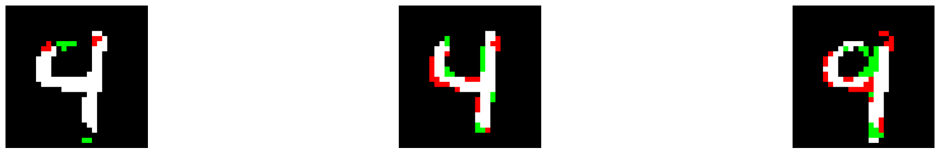
Why feature-based explanations for k-NNs? Traditionally, k-NN models are considered "self-interpretable" because they identify a subset of training data that determines a new input's classification [49]. However, this view is overly simplistic, as interpretability depends on whether individual instances and their features are understandable [44,49]. In high-dimensional settings, the k nearest neighbors may already be too complex for direct human interpretation. Similar challenges arise in other classifiers like decision trees, often viewed as "self-interpretable". This has spurred research into concise, feature-based explanations—such as abductive and counterfactual ones [36]. As the next example shows, applying this approach to NN classification yields meaningful insights.

Example 1. *Consider a* 1*-NN classifier trained on a subset of the MNIST dataset containing digits 4 and 9. The test image in Figure 1a is correctly classified as a 4 based on its NN in Figure 1b, while its closest counterfactual, shown in Figure 1c, is classified as a 9 due to its NN in Figure 1d. The explanation for why the test image is not classified as a 9, highlighted in Figure 1e, identifies 13 pixels that correspond to key differences in the digit's structure. This counterfactual explanation reveals the minimal changes, among the dataset's 784 features, needed to alter the classification.* □

Context. The algorithmic aspects of computing abductive and counterfactual explanations for ML models have garnered significant attention in recent years [4–6,9–11,15,17,29,32–34,37–39,45–47,55,57]. These efforts have explored the computational cost of generating explanations across various ML models, including decision trees, binary decision diagrams, Bayesian networks, neural networks, and graph-based classifiers. Surprisingly, despite the foundational importance of k-NN in machine learning and data management, the literature on explainability for k-NN classifiers remains sparse. The few existing works primarily adopt operational approaches, leveraging mixed integer programming and constraint programming techniques to solve relevant explainability problems [24,42]. However, the theoretical complexity of computing explanations for k-NN models remains an unexplored question, which constitutes a critical gap in the field. In particular, we do not know which of these problems are computationally intractable, and thus can only be approached by applying modern solvers technology for satisfiability or integer programming problems.

Our contributions We perform an analysis of the complexity of checking and computing explanations for k-NN classifiers in two common settings: (a) the *continuous* setting, where points are vectors of real numbers, and the distance is based on the ℓ_p-norm for some integer $p > 0$, and (b) the *discrete* setting, where points are Boolean vectors, and the Hamming distance is used.

(a) Test image. (b) Nearest neighbor of (a). (c) Closest counterfactual. (d) Nearest neighbor of (c).



(e) Diff. map between (a) and (c). (f) Diff. map between (a) and (b). (g) Diff. map between (c) and (d).

Figure 1: Illustration of a counterfactual explanation for an image of digit 4 in the binarized MNIST dataset, which after changing 13 pixels is classified as a 9.

- We start by studying the complexity of checking for the existence of sufficient reasons of a certain size. We show that this problem is NP-hard in both the discrete and continuous settings. In the latter case, hardness holds for every distance ℓ_p, where p is a positive integer.
- We next examine the continuous setting to assess tractability for the remaining problems. Focusing on the ℓ_2 and ℓ_1 distances, two of the most commonly used metrics in practice, we show that tractability heavily depends on the chosen metric.
  - For the ℓ_2-distance, several positive results are established. First, we prove that the problem of checking the existence of a counterfactual explanation at a certain distance is tractable. Furthermore, if such a counterfactual explanation exists, at least one can be computed in polynomial time. Next, we examine the problem of checking if a subset of components from an input $\bar{x}$ constitutes a sufficient reason and prove that this problem is also tractable. Consequently, a *minimal* sufficient reason for $\bar{x}$—where minimality refers to set containment—can always be computed in polynomial time.
  - We then examine the ℓ_1-distance and show that not all positive results derived for the ℓ_2-distance carry over. Specifically, checking the existence of a counterfactual explanation at a certain distance becomes NP-hard in this case. However, for a k-NN classifier with $k = 1$, a *minimal* sufficient reason for input $\bar{x}$ can still be computed in polynomial time.
- Afterwards, we consider the discrete setting and show that it resembles the behavior observed for the ℓ_1-distance in the continuous setting. Specifically: (a) checking the existence of a counterfactual explanation at a certain distance is NP-hard, and (b) for a k-NN classifier with $k = 1$, a *minimal* sufficient reason for an input $\bar{x}$ can be computed in polynomial time. However, this property does not generalize beyond $k = 1$. In particular, even checking if a subset of components of an input $\bar{x}$ is a sufficient reason for a 3-NN classifier is NP-hard. Our results also imply hardness for computing minimal sufficient reasons, in the case $k \geq 3$.
- We present a preliminary analysis of computing explanations for k-NN when $k = 1$, addressing both polynomial-time and NP-hard problems. For the latter, we employ two standard approaches: Integer Quadratic Programming (IQP) and SAT-solving (SAT). While Mixed Integer Programming (MIP) for counterfactual explanations was previously explored [16], our SAT encoding is novel and utilizes a recent solver with native support for cardinality constraints [50]. Interestingly, even for NP-hard problems, explanations for datasets with hundreds of features and thousands of points can be computed in under two minutes.

Organization of the paper Basic definitions are provided in Section 2, and the explainability problems are discussed in Section 3. Negative results concerning minimum sufficient reasons are presented in Section 4. Additional results for the continuous setting, based on the ℓ_2-distance, are covered in Section 5, while those for the ℓ_1-distance are detailed in Section 6. Results for the discrete setting appear in Section 7. Experimental results are discussed in Section 8, while final remarks are included in Section 9.

2 Definitions

Basics We consider pairs of the form (M, D), called *metric space families*, in which M is a set and $D = \{d_n \mid n > 0\}$ satisfies that $d_n : M^n \times M^n \to \mathbb{R}$ is a metric (often referred as distance) on M^n, for every $n > 0$. Elements in M^n are called vectors, and are typically denoted as $\bar{x}, \bar{y}, \bar{z}$. For $i \in \{1, \dots, n\}$, we write $\bar{x}[i]$ to denote the ith component of vector $\bar{x}$.

Metric spaces studied in the paper In this article, we focus on two particular cases for the metric space families of the form (M, D):

- *Continuous case:* Here $M = \mathbb{R}$ and $D = \{d_n \mid n > 0\}$ satisfies that there exists an integer $p > 0$ such that the distance d_n is the one based on the ℓ_p-norm over $\mathbb{R}^n$, for every $n > 0$. In this particular case, we denote D as D_p.
- *Discrete case:* Here $M = \{0, 1\}$ and $D = \{d_n \mid n > 0\}$ satisfies that d_n is the Hamming distance on $\{0, 1\}^n$. That is, if $\bar{x}, \bar{y} \in \{0, 1\}^n$, then $d_n(\bar{x}, \bar{y})$ is the number of components $i \in \{1, \dots, n\}$ for which $\bar{x}[i] \neq \bar{y}[i]$. In this case, we denote D as D_H.

Nearest neighbor classification We fix a metric space family (M, D) as defined above. Let k be a fixed odd integer. Consider two subsets S^+ and S^- of M^n, for $n > 0$, where vectors in S^+ represent *positive examples*, and vectors in S^- represent *negative examples*. For the pair (M, D), we aim to construct a *k-Nearest Neighbor (k-NN) classification function*

$$f^k_{S^+,S^-} : M^n \to \{0, 1\},$$

such that $f^k_{S^+,S^-}(\bar{x}) = 1$ if and only if the majority of the k closest points to $\bar{x}$ in $S^+ \cup S^-$ are positive. However, the set of k closest points may not always be uniquely defined, as multiple points can have the same distance from $\bar{x}$. To address this, we define $f^k_{S^+,S^-}(\bar{x}) = 1$ if and only if there is a subset $T \subseteq S^+ \cup S^-$ of size k such that the majority of points of T belong to S^+ and $d_n(\bar{x}, \bar{y}) \leq d_n(\bar{x}, \bar{z})$ for all $\bar{y} \in T$ and $\bar{z} \in (S^+ \cup S^-) \setminus T$. This approach is sometimes referred to as an *optimistic view* of k-NN classification, as it favors sets that classify $\bar{x}$ as positive when there is ambiguity in the selection of k closest points [16]. In some proofs we use the following characterization of the *optimistic* k-NN classification function, which is immediate from the definition:

Proposition 1. *(a) We have $f^k_{S^+,S^-}(\bar{x}) = 1$ if and only if there exist $A \subseteq S^+$ of size $(k+1)/2$ and $B \subseteq S^-$ of size at most $(k-1)/2$ such that $d_n(\bar{x}, \bar{a}) \leq d_n(\bar{x}, \bar{c})$ for every $\bar{a} \in A$ and $\bar{c} \in S^- \setminus B$.*

(b) We have $f^k_{S^+,S^-}(\bar{x}) = 0$ if and only if there exist $A \subseteq S^-$ of size $(k+1)/2$ and $B \subseteq S^+$ of size at most $(k-1)/2$ such that $d_n(\bar{x}, \bar{a}) < d_n(\bar{x}, \bar{c})$ for every $\bar{a} \in A$ and $\bar{c} \in S^+ \setminus B$.

3 Problems

3.1 Decision problems

We consider a metric space family (M, D) with $D = \{d_n \mid n > 0\}$. We present the different sorts of explanation studied in this paper and their associated decision problems.

Abductive explanations Consider an input vector $\bar{x} \in M^n$. The goal in this case is to find a set X of components over $\{1, \ldots, n\}$ that suffice to explain the output of the k-NN classification function $f^k_{S^+,S^-}$ on $\bar{x}$. Intuitively, this means that every input vector $\bar{y}$ that coincides with $\bar{x}$ over the components in X is classified in the same way by $f^k_{S^+,S^-}$. We formalize these ideas next using the well-known notion of *sufficient reason*.

Fix an odd integer $k \geq 1$. Consider then two sets $S^+, S^- \subseteq M^n$ and an input vector $\bar{x} \in M^n$. Let $X \subseteq \{1, \ldots, n\}$. We call X a *sufficient reason for $\bar{x}$ with respect to $f^k_{S^+,S^-}$*, if

$$f^k_{S^+,S^-}(\bar{x}) = f^k_{S^+,S^-}(\bar{y}), \quad \text{for every } \bar{y} \in M^n \text{ that satisfies } \bar{x}[i] = \bar{y}[i], \text{ for each } i \in X.$$

The most basic decision problem in this case is verifying if an $X \subseteq \{1, \ldots, n\}$ is in fact a sufficient reason for $\bar{x}$. This leads to the following problem.

PROBLEM :	k-Check Sufficient Reason(M, D)
INPUT :	Two sets $S^+, S^- \subseteq M^n$, a vector $\bar{x} \in M^n$, an $X \subseteq \{1, \ldots, n\}$
OUTPUT :	Yes, if X is a sufficient reason for $\bar{x}$ with respect to $f^k_{S^+,S^-}$

Not all sufficient reasons are equally informative. For instance, $X = \{1, \ldots, n\}$ is always a sufficient reason for $\bar{x}$, but arguably a very uninformative one. It is then natural to look for *minimum* sufficient reasons, that is, sufficient reasons that are as small as possible in terms of their cardinality. This is formalized by the next decision problem.

PROBLEM :	k-Minimum Sufficient Reason(M, D)
INPUT :	Two sets $S^+, S^- \subseteq M^n$, a vector $\bar{x} \in M^n$, an integer $\ell > 0$
OUTPUT :	Yes, if there is a sufficient reason X for $\bar{x}$ w.r.t. $f^k_{S^+,S^-}$ with $\|X\| \leq \ell$

When the problem of checking minimum sufficient reasons is computationally hard, one might be satisfied with finding a *minimal* one, i.e., one that does not properly contain another sufficient reason. Formally, if X is a sufficient reason for $\bar{x}$ with respect to $f^k_{S^+,S^-}$, then X is *minimal* if there is no sufficient reason Y for $\bar{x}$ with respect to $f^k_{S^+,S^-}$ that satisfies $Y \subsetneq X$. Clearly, every minimum sufficient reason is also minimal, but the converse does not hold in general as shown next.

Example 2. *We consider the discrete setting. Assume that $S^+ = \{(0, 1, 1), (1, 0, 1), (1, 1, 1)\}$ and $S^- = \{0, 1\}^3 \setminus S^+$. It is easy to see that both sets $\{1, 2\}$ and $\{3\}$ of components are sufficient reasons for $\bar{x} = (0, 0, 0)$ with respect to $f^k_{S^+,S^-}$, for every odd integer $k \geq 1$. This is because every vector $\bar{y}$ for which $\bar{y}[1] = \bar{y}[2] = 0$, or for which $\bar{y}[3] = 0$, belongs to S^-, and thus $f^k_{S^+,S^-}(\bar{x}) = f^k_{S^+,S^-}(\bar{y}) = 0$. Moreover, neither $\{1\}$ nor $\{2\}$ nor $\emptyset$ are sufficient reasons for $\bar{x}$. Hence, $\{1, 2\}$ and $\{3\}$ are minimal sufficient reasons for $\bar{x} = (0, 0, 0)$, but only $\{3\}$ is a minimum one.* □

This motivates our next decision problem.

PROBLEM :	k-Minimal Sufficient Reason(M, D)
INPUT :	Two sets $S^+, S^- \subseteq M^n$, a vector $\bar{x} \in M^n$, an $X \subseteq \{1, \ldots, n\}$
OUTPUT :	Yes, if X is a minimal sufficient reason for $\bar{x}$ w.r.t. $f^k_{S^+,S^-}$

It is easy to observe that a greedy strategy turns a polynomial time algorithm for k-Check Sufficient Reason into a polynomial time algorithm for k-Minimal Sufficient Reason.

Proposition 2. *For any k, M, D, the k-*Minimal Sufficient Reason*(M, D) problem reduces in polynomial time to k-*Check Sufficient Reason*(M, D).*

Proof. If a set is a sufficient reason, then all its supersets are. Hence, to decide if $X \subseteq \{1, \ldots, n\}$ is a minimal sufficient reason, it suffices to check if X is a sufficient reason, and then check, for each subset $X \setminus \{i\}$ obtained by removing one element $i \in X$, that $X \setminus \{i\}$ is not a sufficient reason. □

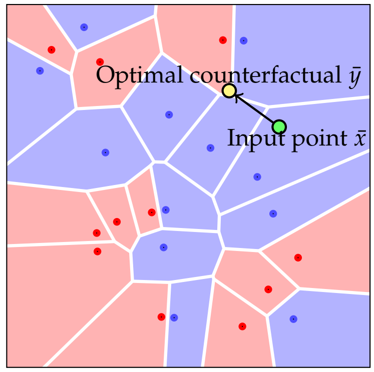


Figure 2: Illustration of an minimum distance counterfactual explanation over $\mathbb{R}^2$ in the ℓ_2 metric. Blue (red) areas are classified negatively (positively).

Counterfactual explanations These are explanations that aim to find what should be changed from an input vector $\bar{x}$ in order to obtain a different classification outcome. Typically, one aims to find counterfactual explanations that are not "too far" from $\bar{x}$, which is formalized by saying that the distance between $\bar{x}$ and its counterfactual explanation is bounded.

Fix an odd integer $k \geq 1$. Given sets $S^+, S^- \subseteq M^n$ and $\bar{x} \in M^n$, a *counterfactual explanation for $\bar{x}$ with respect to $f^k_{S^+,S^-}$* is a vector $\bar{y} \in M^n$ with $f^k_{S^+,S^-}(\bar{x}) \neq f^k_{S^+,S^-}(\bar{y})$. We look for counterfactual explanations that are close to the vector $\bar{x}$. This leads to the following decision problem.

PROBLEM :	k-Counterfactual Explanation(M, D)
INPUT :	Two sets $S^+, S^- \subseteq M^n$, a vector $\bar{x} \in M^n$, a rational $\ell > 0$
OUTPUT :	Yes, if there is a counterfactual explanation $\bar{y}$ for $\bar{x}$ with respect to $f^k_{S^+,S^-}$ such that $d_n(\bar{x}, \bar{y}) \leq \ell$

Thus, Counterfactual Explanation asks whether it is possible to find a vector $\bar{y}$ that is relatively close to $\bar{x}$ and that is classified differently than $\bar{x}$ under $f^k_{S^+,S^-}$. For instance, in the discrete case this asks if it is possible to "flip" the classification of $\bar{x}$ by "flipping" at most ℓ of its components. Figure 2 illustrates how counterfactuals look in the continuous setting under the ℓ_2-distance.

3.2 Computation problems

For simplicity, we focus our complexity analysis on the decision problems introduced above. However, in the context of explainable AI, it is, of course, more important to compute an optimal explanation (if one exists). Our study, however, also sheds light on the computational problem. In fact, as shown in the explainability literature, the hardness of a decision problem often implies hardness for its associated computation problem [5]. Conversely, the tractability of a decision problem often implies that the associated computation problem can be solved in polynomial time. In this paper, we show that all our tractability results extend from decision to computation.

4 Minimum Sufficient Reasons

In this section, we show that k-Minimum Sufficient Reason is NP-hard for both the continuous and the discrete setting, for every odd integer $k \geq 1$. In the continuous case, hardness holds regardless of the norm being used.

Theorem 1. *The following statements hold:*

1. *The problem k-*Minimum Sufficient Reason*$(\mathbb{R}, D_p)$ is* NP-*hard, for every fixed odd integer $k \geq 1$ and integer $p > 0$.*
2. *The problem k-*Minimum Sufficient Reason*$(\{0,1\}, D_H)$ is* NP-*hard, for every fixed odd integer $k \geq 1$.*

Proof. We start by proving (1), and then derive (2) by a modification of the proof. We reduce from the well-known NP-complete *Vertex Cover* problem: given an undirected graph $G = (V, E)$, and an integer $\ell \geq 0$, check whether there is a *vertex cover* C in G of size $|C| \leq \ell$. Recall that a vertex cover is a subset of nodes $C \subseteq V$ such that every edge in E has an endpoint in C.

Given an instance of Vertex Cover, we construct an instance of the problem k-Minimum Sufficient Reason$(\mathbb{R}, D_p)$ as follows. Assume that $V = \{1, \ldots, n\}$ and $E = \{e_1, \ldots, e_m\}$, for $n, m \geq 1$. Suppose that n is the vector dimension, and take $\bar{x} = (0, \ldots, 0) \in \mathbb{R}^n$. Choose $(k+1)/2$ numbers such that $1/2 > \varepsilon_1 > \cdots > \varepsilon_{(k+1)/2} > 0$. For each $j \in \{1, \ldots, m\}$, and $h \in \{1, \ldots, (k+1)/2\}$, we define the vector $\bar{y}_{j,h} \in \mathbb{R}^n$ such that $\bar{y}_{j,h}[i] = 1 + \varepsilon_h$ if e_j is incident to the vertex i, and $\bar{y}_{j,h}[i] = 0$, otherwise. We also define:

$$S^- = \{\bar{y}_{j,h} \mid j \in \{1, \ldots, m\}, h \in \{1, \ldots, (k+1)/2\}\}.$$

For $\bar{y}_{j,h}$, we denote by $\bar{y}^1_{j,h}$ and $\bar{y}^2_{j,h}$, the vectors obtained from $\bar{y}_{j,h}$ by changing the first and second component $1 + \varepsilon_h$, respectively, by ε_h, and keeping the remaining vector components unchanged. We finally define:

$$S^+ = \bigcup_{j,h} \{\bar{y}^1_{j,h}, \bar{y}^2_{j,h}\}.$$

Note that $f^k_{S^+,S^-}(\bar{x}) = 1$, as for every $\bar{y}^a_{j,(k+1)/2} \in S^+$ and $\bar{y}_{j',h} \in S^-$:

$$\|\bar{y}^a_{j,(k+1)/2}\|^p_p = \varepsilon^p_{(k+1)/2} + (1 + \varepsilon_{(k+1)/2})^p < 2(1+\varepsilon_h)^p = \|\bar{y}_{j',h}\|^p_p.$$

We claim that there is a vertex cover C with $|C| \leq \ell$ if and only if there is a sufficient reason X for $\bar{x}$ with respect to $f^k_{S^+,S^-}$ such that $|X| \leq \ell$. Suppose first there is such a vertex cover $C \subseteq \{1, \ldots, n\}$. We show that C is a sufficient reason. Let $\bar{z} \in \mathbb{R}^n$ be an arbitrary vector such that $\bar{z}[i] = \bar{x}[i] = 0$, for all $i \in C$. We show that there is an inyective function $g : S^- \to S^+$ such that for every $\bar{y} \in S^-$, we have $\|\bar{z} - \bar{y}\|^p_p > \|\bar{z} - g(\bar{y})\|^p_p$, and hence $f^k_{S^+,S^-}(\bar{z}) = 1$ as required. Let $\bar{y}_{j,h} \in S^-$. Since C is a vertex cover, one of the endpoints of e_j is in C and then there is $i \in C$ such that $\bar{y}_{j,h}[i] = 1 + \varepsilon_h$. Pick one such i, and define $g(\bar{y}_{j,h}) \in S^+$ as the vector resulting from $\bar{y}_{j,h}$ by changing $\bar{y}_{j,h}[i]$ to ε_h. The function g is inyective:

- for $\bar{y}_{j,h}, \bar{y}_{j',h'} \in S^-$ with $j \neq j'$, we have $g(\bar{y}_{j,h}) \neq g(\bar{y}_{j',h'})$ as their non-zero components differ; and
- for $\bar{y}_{j,h}, \bar{y}_{j,h'} \in S^-$ with $h \neq h'$, we have $g(\bar{y}_{j,h}) \neq g(\bar{y}_{j,h'})$ as $\varepsilon_h \neq \varepsilon_{h'}$.

Since $\bar{y}_{j,h}$ and $g(\bar{y}_{j,h})$ only differ in one component $i \in C$, we have

$$\|\bar{z} - \bar{y}_{j,h}\|^p_p > \|\bar{z} - g(\bar{y}_{j,h})\|^p_p \iff |\bar{z}[i] - (1+\varepsilon_h)|^p > |\bar{z}[i] - \varepsilon_h|^p \iff (1+\varepsilon_h)^p > \varepsilon^p_h,$$

and then g satisfied the required conditions.

Assume now that $X \subseteq \{1, \ldots, n\}$ is a sufficient reason with $|X| \leq \ell$. We show that X is a vertex cover of G. By contradiction, suppose there is an edge $e_j \in E$ whose endpoints are not in X. Then the vector $\bar{y}_{j,1}$ satisfies that $\bar{y}_{j,1}[i] = \bar{x}[i] = 0$ for all $i \in X$. We claim that $f^k_{S^+,S^-}(\bar{y}_{j,1}) = 0$, which is a contradiction. This follows from the fact that $\|\bar{y}_{j,1} - \bar{y}_{j,h}\|^p_p < \|\bar{y}_{j,1} - \bar{y}^a_{j',h'}\|^p_p$, for every $h \in \{1, \ldots, (k+1)/2\}$ and $\bar{y}^a_{j',h'} \in S^+$. Indeed, since $\bar{y}^a_{j',h'}$ contains at most one component with value $1 + \varepsilon_{h'}$, and $\bar{y}_{j,1}$ has two components with value $1 + \varepsilon_1$, we have:

$$\|\bar{y}_{j,1} - \bar{y}^a_{j',h'}\|^p_p \geq (1 + \varepsilon_1 - \varepsilon_{h'})^p \geq 1 \geq 2\left(\frac{1}{2}\right)^p > 2(\varepsilon_1 - \varepsilon_h)^p = \|\bar{y}_{j,1} - \bar{y}_{j,h}\|^p_p.$$

This finishes the proof of (1).

We prove the remaining discrete case (2). We first consider the case $k = 1$. The proof follows the same strategy than in the continuous case (1). Again we reduce from the *Vertex Cover* problem: given an undirected graph $G = (V, E)$, and an integer $\ell \geq 0$, check whether there is a *vertex cover* C in G of size $|C| \leq \ell$.

Given an instance of Vertex Cover, we construct an instance of the problem 1-Minimum Sufficient Reason($\{0, 1\}, D_H$) as follows. Assume that $V = \{1, \ldots, n\}$ and $E = \{e_1, \ldots, e_m\}$, for $n, m \geq 1$. Suppose that n is the vector dimension, and take $\bar{x} = (0, \ldots, 0) \in \{0, 1\}^n$. For each $j \in \{1, \ldots, m\}$, we define the vector $\bar{y}_j \in \{0, 1\}^n$ such that $\bar{y}_j[i] = 1$ if e_j is incident to the vertex i, and $\bar{y}_j[i] = 0$, otherwise. We define $S^- = \{\bar{y}_j \mid j \in \{1, \ldots, m\}\}$. For $\bar{y}_j$, we denote by $\bar{y}_j^1$ and $\bar{y}_j^2$, the vectors obtained from $\bar{y}_j$ by flipping the first and second component with value 1, respectively, to 0, and keeping the remaining vector components unchanged. We finally define $S^+ = \bigcup_j \{\bar{y}_j^1, \bar{y}_j^2\}$. Note that $f^1_{S^+,S^-}(\bar{x}) = 1$, as $d_H(\bar{x}, \bar{y}) = 1$ for every $\bar{y} \in S^+$, while $d_H(\bar{x}, \bar{y}) = 2$ for every $\bar{y} \in S^-$.

We claim that there is a vertex cover C with $|C| \leq \ell$ if and only if there is a sufficient reason X for $\bar{x}$ with respect to $f^1_{S^+,S^-}$ such that $|X| \leq \ell$. Suppose first there is such a vertex cover $C \subseteq \{1, \ldots, n\}$. We show that C is a sufficient reason. Let $\bar{z} \in \{0, 1\}^n$ be an arbitrary vector such that $\bar{z}[i] = \bar{x}[i] = 0$, for all $i \in C$. We show that for every $\bar{y} \in S^-$, there exists $\bar{y}' \in S^+$, such that $d_H(\bar{z}, \bar{y}) > d_H(\bar{z}, \bar{y}')$, and hence $f^1_{S^+,S^-}(\bar{z}) = 1$ as required. Let $\bar{y}_j \in S^-$. Since C is a vertex cover, one of the endpoints of e_j is in C and then there is $i \in C$ such that $\bar{y}_j[i] = 1$. Pick one such i, and define $\bar{y}'_j \in S^+$ as the vector resulting from $\bar{y}_j$ by flipping $\bar{y}_j[i]$ to 0. Since $\bar{y}_j$ and $\bar{y}'_j$ only differ in one component $i \in C$, we have

$$d_H(\bar{z}, \bar{y}_j) > d_H(\bar{z}, \bar{y}'_j) \iff |\bar{z}[i] - \bar{y}_j[i]| > |\bar{z}[i] - \bar{y}'_j[i]| \iff 1 > 0$$

and then the condition holds.

Assume now that $X \subseteq \{1, \ldots, n\}$ is a sufficient reason with $|X| \leq \ell$. We show that X is a vertex cover of G. By contradiction, suppose there is an edge $e_j \in E$ whose endpoints are not in X. Then the vector $\bar{y}_j$ satisfies that $\bar{y}_j[i] = \bar{x}[i] = 0$ for all $i \in X$. As $\bar{y}_j \in S^-$, it follows that $f^1_{S^+,S^-}(\bar{y}_j) = 0$, which is a contradiction. The hardness of the case $k \geq 3$ follows directly from the proof of Theorem 5 as for $k \geq 3$, the problem k-Check Sufficient Reason($\{0, 1\}, D_H$) is hard, even when the input subset of components is $X = \emptyset$. Hardness for k-Minimum Sufficient Reason($\{0, 1\}, D_H$) follows directly by setting the input threshold $\ell = 0$.

□

This resolves the complexity of one of our main problems across all the settings examined in the paper. In the next two sections, we address the remaining problems in the continuous setting, specifically exploring how the choice of metric affects their tractability. To this end, we focus on two of the most commonly studied metrics: ℓ_2 and ℓ_1.

5 The Continuous Setting Based on the ℓ_2-distance

It turns out that in case of the ℓ_2-norm, all mentioned problems, apart from k-Minimum Sufficient Reason, are tractable. The main reason is that in the case of the ℓ_2-norm, an inequality of the form "the point $\bar{x}$ is closer to the point $\bar{a}$ than to the point $\bar{c}$" is the *linear* inequality in $\bar{x}$ given by $(\bar{a} - \bar{c})^\top \bar{x} \geq \frac{1}{2}(\bar{a} - \bar{c})^\top(\bar{a} + \bar{c})$. By Proposition 1, this gives a representation of the set $\{\bar{x} \in \mathbb{R}^n \mid f_{S^+,S^-}(\bar{x}) = 1\}$ as a union of at most $|S^+ \cup S^-|^{2k}$ many polyhedra, a polynomial in the input. These polyhedra are explicitly given, as we can describe them by a system of linear inequalities in polynomial time. Analogously, the set $\{\bar{x} \in \mathbb{R}^n \mid f_{S^+,S^-}(\bar{x}) = 0\}$ is a union of polynomially many "open polyhedra", that is, sets of solutions to a system of *strict* linear inequalities.

Abductive explanations We start by showing tractability of k-Check Sufficient Reason. By Proposition 2, this implies tractability of k-Minimal Sufficient Reason for the ℓ_2-norm, which in turn implies that minimal sufficient reasons can be computed in polynomial time in this case.

Proposition 3. *The problem k-Check Sufficient Reason$(\mathbb{R}, D_2)$, and hence also k-Minimal Sufficient Reason$(\mathbb{R}, D_2)$, can be solved in polynomial time for every fixed odd integer $k \geq 1$.*

Proof. Assume first that $f^k_{S^+,S^-}(\bar{x}) = 0$. Let $X \subseteq \{1, \dots, n\}$ and consider the affine subspace $U(X, \bar{x}) := \{y \in \mathbb{R}^n \mid \bar{x}[i] = y[i]\}$. Then X is *not* a sufficient reason for $\bar{x}$ if and only if $U(X, \bar{x})$ intersects the set $\{\bar{y} \in \mathbb{R}^n \mid f_{S^+,S^-}(\bar{y}) = 1\}$. By Proposition 1, this set is a union of polynomially many polyhedra. It remains to check if our affine subspace intersects one of these polyhedra. The ntersection of an affine subspace and a polyhedron is a polyhedron, and checking emptiness of a polyhedron is equivalent to linear programming which thus can be done in polynomial time [54].

In the case $f^k_{S^+,S^-}(\bar{x}) = 1$, we have to check, whether our affine subspace intersects the set $\{\bar{y} \in \mathbb{R}^n \mid f_{S^+,S^-}(\bar{y}) = 0\}$. This time, by Proposition 1, this set is a union of sets of solutions to systems of strict linear inequalities. The same argument as in the previous case reduces our problem to the emptiness problem for an intersection of an affine subspace with an open polyhedron. The latter is just the feasibility problem for systems of linear equalities and strict linear inequalities, which can be reduced to linear programming (with non-strict inequalities), solving our problem in polynomial time. Namely, let S be a system of linear equalities and strict linear inequalities. Consider a system $\widehat{S}$ of non-strict linear inequalities over variables of S and a new variable ε, obtained by turning every strict inequality $l > 0$ of S into a non-strict inequality $l \geq \varepsilon$. Feasibility of S is equivalent to existence of a feasible solution to $\widehat{S}$ with positive ε. To find out if the latter is true, it is enough to find the optimal solution to the problem of maximizing ε subject to $\widehat{S}$. ◻

As a corollary, we obtain the following:

Corollary 1. *Consider the setting $(\mathbb{R}, D_2)$ and let $k \geq 1$ be any odd integer. There is a polynomial time algorithm that, given sets $S^+, S^- \subseteq \mathbb{R}^n$ and a vector $\bar{x} \in \mathbb{R}^n$, it computes a minimal sufficient reason $\bar{y}$ for $\bar{x}$ with respect to $f^k_{S^+,S^-}$.*

Proof. This reduces to finding a point in the intersection of the ball $B_R(\bar{x})$ and an open polyhedron P, if this intersection is non-empty. If it is non-empty, then $\widehat{P}$ intersects the interior of $B_R(\bar{x})$. We can find a point in this intersection by minimizing the quadratic form $q(\bar{y}) = \|\bar{x} - \bar{y}\|_2^2 \to \min$ subject to $\widehat{P}$. This point $\bar{y}$ will belong to the border of $\widehat{P}$. That is, some inequalities, defining $\widehat{P}$, turn into equalities on $\bar{y}$. We now need to find a direction that points to the interior of $\widehat{P}$ from this point. More precisely, our task is to find a vector β such that $\langle \alpha, \beta \rangle > 0$ for every inequality $\langle \alpha, \bar{y} \rangle \geq c$ which turns into equality on $\bar{y}$. After such β is found, we just need to move from $\bar{y}$ along β a little bit so that we still inside $B_R(\bar{x})$. Finding such β is reducible to finding a solution to a system of strict linear inequalities. In turn, the latter can be reduced in polynomial time to linear programming as explained in the proof of Proposition 3. ◻

Counterfactual explanations Tractability of the k-Counterfactual Explanation problem is proved similarly to k-Check Sufficient Reason, but this time we use polynomial-time solvability of convex quadratic programming [41].

Theorem 2. *The problem k-Counterfactual Explanation$(\mathbb{R}, D_2)$ can be solved in polynomial time for every fixed odd integer $k \geq 1$.*

Proof. In the k-Counterfactual Explanation$(\mathbb{R}, D_2)$ problem, given $\bar{x} \in \mathbb{R}^n$ and $\ell > 0$, the goal is to check if the ball $B_\ell(\bar{x}) = \{\bar{y} \in \mathbb{R}^n \mid \|\bar{y} - \bar{x}\|_2 \leq \ell\}$ contains a vector with a different $f^k_{S^+,S^-}$ value than $\bar{x}$. If $f^k_{S^+,S^-}(\bar{x}) = 0$, this reduces to checking whether $B_\ell(\bar{x})$ intersects the set $\{\bar{y} \in \mathbb{R}^n \mid f^k_{S^+,S^-}(\bar{y}) \neq 0\}$, which, by Proposition 1, is a union of polynomially many polyhedra. Thus, the problem reduces to determining whether $B_\ell(\bar{x})$ intersects a given polyhedron P. This can be solved via convex quadratic programming, which minimizes a positive definite quadratic form under a system of non-strict linear inequalities and is solvable in polynomial time due to Kozlov, Tarasov, and Khachiyan [41]. Specifically, we minimize $q(\bar{y}) = \|\bar{x} - \bar{y}\|_2^2$ subject to constraints defining P. The answer to k-Counterfactual Explanation$(\mathbb{R}, D_2)$ is Yes if and only if the minimum value is at most ℓ^2, with the optimal solution providing the counterfactual explanation.

Similarly, when $f^k_{S^+,S^-}(\bar{x}) = 1$, the problem reduces to checking whether the ball $B_\ell(\bar{x})$ intersects a given open polyhedron, defined as the solution set to a system of strict linear inequalities. The argument from the previous paragraph requires modification because the algorithm in [41] assumes non-strict inequalities in the constraints. To address this, we first check whether P is empty, reducing the problem to linear programming as described in the proof of Proposition 3. If P is non-empty (since otherwise, there is nothing left to check), we construct a polyhedron $\widehat{P}$ by converting all strict inequalities of P into non-strict ones. Note that P corresponds to the interior of $\widehat{P}$. We claim that P intersects $B_\ell(\bar{x})$ if and only if $\widehat{P}$ intersects the *interior* of $B_\ell(\bar{x})$. The latter can be reduced to a problem of minimizing a convex quadratic objective subject to (closed) polyhedron, which in turn can be solved in polynomial time with the techniques by Kozlov, Tarasov, and Khachiyan [41]. Indeed, consider the problem of minimizing $q(\bar{y}) = \|\bar{x} - \bar{y}\|_2^2$ subject to $\bar{y} \in \widehat{P}$ (which is a system of non-strict inequalities, as required in the algorithm). The answer is Yes if and only if the optimum is strictly less than ℓ^2.

We now establish the claim. First, assume that P intersects $B_\ell(\bar{x})$. We show that P (and hence $\widehat{P}$) also intersects the interior of $B_\ell(\bar{x})$. If P has a point on the boundary of the ball, it must also have a point inside, as P is open and any boundary point of the ball has interior points arbitrarily close to it. Now assume that $\widehat{P}$ intersects the interior of $B_\ell(\bar{x})$. Since P is the interior of $\widehat{P}$, it is non-empty, and $\widehat{P}$ is a full-dimensional polyhedron. By Proposition 2.1.8 in [31], every point in a full-dimensional polyhedron has interior points arbitrarily close to it. Thus, if $\widehat{P}$ intersects the interior of the ball, P, as the interior of $\widehat{P}$, must also intersect the interior of the ball. □

By extending the techniques used in the proof of Theorem 2, we can conclude that in the continuous setting, a counterfactual explanation can be computed in polynomial time, provided one exists, when ℓ_2-distances are used.

Corollary 2. *Consider the setting $(\mathbb{R}, D_2)$ and let $k \geq 1$ be any odd integer. There is a polynomial time algorithm that, given sets $S^+, S^- \subseteq \mathbb{R}^n$, vector $\bar{x} \in \mathbb{R}^n$, and rational $\ell > 0$, it computes a counterfactual explanation for $\bar{x}$ with respect to $f^k_{S^+,S^-}$ at distance at most ℓ in case there exists one.*

6 The Continuous Setting Based on the ℓ_1-distance

We first show that the positive results for counterfactual explanations under the ℓ_2-norm do not extend to the ℓ_1-norm. For abductive explanations, we demonstrate that some favorable properties of the ℓ_2-norm remain, allowing minimal sufficient reasons to be computed efficiently when $k = 1$.

Counterfactual explanations We show that the k-Counterfactual Explanation problem for the ℓ_1-distance is NP-hard even when S^+ and S^- are of minimal non-degenerate size.

Theorem 3. *For every odd integer $k \geq 1$, the problem k-Counterfactual Explanation$(\mathbb{R}, D_1)$ is* NP*-complete even when $|S^+| = |S^-| = (k+1)/2$.*

Proof. The NP upper bound is straightforward: if a counter-factual explanation exists, it can be found as a solution to a polynomially bounded linear program. We now show the lower bound. In the appendix we show that it suffices to establish NP-hardness for $k = 1$. We reduce from a variation of the knapsack problem, where the goal is to determine whether at least half the total value of all items can fit into a given knapsack. Specifically, we are given n items, each with a weight w_i and value v_i (positive integers for $i = 1, \ldots, n$), and a positive integer $W > 0$, representing the knapsack's maximum weight capacity. The question is whether there exists a subset $T \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in T} w_i \leq W$ and $\sum_{i \in T} v_i \geq (v_1 + \ldots + v_n)/2$. The hardness of this problem arises from a reduction from the partition problem [43].

We reduce to the 1-Counterfactual Explanation$(\mathbb{R}, D_1)$ problem with $|S^+| = |S^-| = 1$ as follows. The dimension n corresponds to the number of items in the knapsack instance. We set $\bar{x} = \bar{0} \in \mathbb{R}^n$ and the radius

$\ell = W$. The sets $S^+ = \{\bar{g}\}$ and $S^- = \{\bar{h}\}$ are defined as:

$$\bar{g}_i = w_i, \qquad \bar{h}_i = w_i - \gamma \cdot v_i, \qquad i = 1, \ldots, n,$$

where $\gamma = 1/(2 \max_i v_i)$ ensures $\gamma \cdot v_i \le 1/2 < 1$ for all $i = 1, \ldots, n$. Since w_i is a positive integer, we have $0 < \bar{h}_i < \bar{g}_i$, so the interval $[\bar{h}_i, \bar{g}_i]$ lies to the right of 0 and has a length of $\gamma \cdot v_i$. These properties imply $\|\bar{h} - 0\|_1 < \|\bar{g} - 0\|_1$, resulting in $f^1_{S^+,S^-}(\bar{0}) = 0$. We are asked if there exists $\bar{y}$ with $\|y\|_1 \le \ell = W$ such that $f^1_{S^+,S^-}(\bar{y}) = 1$, or, equivalently, $\|\bar{h} - \bar{y}\|_1 \ge \|\bar{g} - \bar{y}\|_1$. We show that the answer is Yes if and only if the original knapsack instance has a solution.

Assume first that the original knapsack instance has a solution. Define a vector $\bar{y} \in \mathbb{R}^n$ by setting $\bar{y}_i = 0$ for items not placed in the knapsack and $\bar{y}_i = \bar{g}_i = w_i$ for items that are placed. Note that $\|\bar{y}\|_1$ equals the total weight of the items in the knapsack, which does not exceed $W = \ell$. We now need to show that $\|\bar{h} - \bar{y}\|_1 \ge \|\bar{g} - \bar{y}\|_1$, or equivalently:

$$\|\bar{h} - \bar{y}\|_1 - \|\bar{g} - \bar{y}\|_1 \;=\; \sum_{i=1}^{n} (|\bar{h}_i - \bar{y}_i| - |\bar{g}_i - \bar{y}_i|) \;\ge\; 0.$$

It is more convenient to express this inequality in the equivalent form:

$$\sum_{i=1}^{n} (|\bar{h}_i - \bar{y}_i| - |\bar{g}_i - \bar{y}_i| + \gamma v_i)/2 \;\ge\; \gamma(v_1 + \ldots + v_n)/2. \tag{1}$$

The term $|\bar{h}_i - \bar{y}_i| - |\bar{g}_i - \bar{y}_i|$ represents the distance from $\bar{y}_i$ to the left endpoint of $[\bar{h}_i, \bar{g}_i]$ minus the distance from $\bar{y}_i$ to the right endpoint of $[\bar{h}_i, \bar{g}_i]$. It is minus the length of the interval to the left of it, and plus the length of the interval to the right of it, and the length of the interval in our case is γv_i. Thus, if $\bar{y}_i = 0$, the left-hand side of (1) contributes 0; if $\bar{y}_i = \bar{g}_i$ (the right endpoint), it contributes γv_i. Therefore, the left-hand side of (1) is the sum of the values of the items in the knapsack, scaled by γ. This proves (1), as it follows directly from the fact that we began with a feasible solution to the original knapsack problem.

We now show the other direction. Assume that there exists $\bar{y} \in \mathbb{R}^n$ such that $\|y\|_1 \le \ell = W$ and $\|\bar{h} - \bar{y}\|_1 \ge \|\bar{g} - \bar{y}\|_1$, with the latter being equivalent to (1). Consider again the quantity $|\bar{h}_i - \bar{y}_i| - |\bar{g}_i - \bar{y}_i|$ as a function of $\bar{y}_i$. To the right of $\bar{g}_i = w_i$ it is constant and is equal to the length of the interval. Hence, without loss of generality, $\bar{y}_i \le \bar{g}_i$ for every $i = 1, \ldots, n$, as otherwise we can decrease $\|y\|_1$ without changing the left-hand side of (1). Likewise, to the left of the interval, the quantity in question is also constant and is equal to the minus of the length of the interval, and the minimal absolute value there is 0. Hence, we may assume that if $\bar{y}_i$ is not in $[\bar{h}_i, \bar{g}_i]$, then $\bar{y}_i = 0$ for every $i = 1, \ldots, n$. Again, if not, $\|y\|_1$ can be decreased without changing the left-hand side of (1).

We now establish that, without loss of generality, we may assume $\bar{y}_i = 0$ or $\bar{y}_i = \bar{g}_i$ for every $i = 1, \ldots, n$. First, suppose there are two different components, $\bar{y}_i$ and $\bar{y}_j$, that lie within their respective intervals but are strictly smaller than their right endpoints. Start decreasing $\bar{y}_i$ and increasing $\bar{y}_j$ at the same rate. The terms involving $\bar{y}_i$ and $\bar{y}_j$ in (1) will begin to decrease and increase, respectively, at twice the same rate, leaving the left-hand side of (1) unchanged. Similarly, $\|\bar{y}\|_1$ remains constant. This process continues until either $\bar{y}_i$ reaches $\bar{h}_i$ or $\bar{y}_j$ reaches $\bar{g}_j$. In the first case, we can further decrease $\bar{y}_i$ to 0 without reducing the left-hand side of (1). In the second case, $\bar{y}_j$, which was strictly within its interval, becomes equal to its right endpoint. In both scenarios, the number of indices i for which $\bar{y}_i \in [\bar{h}_i, \bar{g}_i)$ strictly decreases. We repeat this procedure until at most one index i satisfies $\bar{y}_i \in [\bar{h}_i, \bar{g}_i)$.

Now, if at most one problematic i remains where $\bar{y}_i \in [\bar{h}_i, \bar{g}_i)$, we simply increase $\bar{y}_i$ to $\bar{g}_i$. We claim that the solution remains feasible. Increasing $\bar{y}_i$ raises the left-hand side of (1), so the corresponding inequality is still satisfied. We now explain why $\|\bar{y}\|_1 \le W = \ell$ continues to hold. This follows from the fact that W is an integer and the length of any interval is less than 1. Before the increase, $\|\bar{y}\|_1 = |\bar{y}_1| + \ldots + |\bar{y}_n|$ was at most W. After the increase, the sum becomes integral because $\bar{y}_i = 0$ or $\bar{y}_i = \bar{g}_i = w_i$ for every i. Consequently, the total sum cannot exceed W, as the increase is too small to make the sum reach $W + 1$.

Therefore, it holds that $\bar{y}_i = 0$ or $\bar{y}_i = \bar{g}_i = w_i$ for every $i = 1, \ldots, n$. We place objects with $\bar{y}_i = w_i$ into the knapsack. The condition $|y|_1 \leq \ell = W$ ensures that the sum of the weights of the objects placed in the knapsack does not exceed the capacity. Now, notice that the left-hand side of (1) becomes the sum of γv_i for the objects placed in the knapsack, which guarantees that the total value of the objects placed is at least half of the total value of all objects. ◻

Abductive explanations Next, we observe that the 1-Check Sufficient Reason problem is tractable for the ℓ_1-norm (together with the 1-Minimal Sufficient Reason, by Proposition 2).

Proposition 4. *The problem* 1-Check Sufficient Reason$(\mathbb{R}, D_1)$*, and hence also* 1-Minimal Sufficient Reason$(\mathbb{R}, D_1)$*, is polynomial-time solvable.*

Proof. For a given $S^+, S^- \subseteq \mathbb{R}^n$ $\bar{x} \in \mathbb{R}^n$ and $X \subseteq \{1, \ldots, n\}$, our task is to decide, whether there is a vector, coinciding with $\bar{x}$ on coordinates in X but differing in the value of $f^1_{S^+,S^-}$. We will use the following notation: for a vector $\bar{v} \in \mathbb{R}^n$, we write $\bar{v} = (\bar{v}_1, \bar{v}_2)$, denoting by $\bar{v}_1$ the projection of $\bar{v}$ to coordinates from X and by $\bar{v}_2$ the projection to the remaining coordinates.

In this notation, our task is to see, if there is $\bar{y}_2 \in \mathbb{R}^{[n]\setminus X}$ with the property that $f^1_{S^+,S^-}((\bar{x}_1, \bar{y}_2)) \neq f^1_{S^+,S^-}((\bar{x}_1, \bar{x}_2))$. First, assume that $f^1_{S^+,S^-}((\bar{x}_1, \bar{x}_2)) = 0$. By Proposition 1 for $k = 1$, we are asked if there exists $\bar{y}_2 \in \mathbb{R}^{[n]\setminus X}$ and $\bar{a} \in S^+$ such that:

$$\|(\bar{x}_1, \bar{y}_2) - (\bar{a}_1, \bar{a}_2)\|_1 \leq \|(\bar{x}_1, \bar{y}_2) - (\bar{c}_1, \bar{c}_2)\|_1 \text{ for every } \bar{c} \in S^-. \tag{2}$$

Using linearity of the ℓ_1-norm under concatenation of vectors, we rewrite (2) as follows:

$$\|\bar{x}_1 - \bar{a}_1\|_1 - \|\bar{x}_1 - \bar{c}_1\|_1 \leq \|\bar{y}_2 - \bar{c}_2\|_1 - \|\bar{y}_2 - \bar{a}_2\|_1 \text{ for every } \bar{c} \in S^-. \tag{3}$$

The left-hand side of (3) does not depend on $\bar{y}_2$. The right-hand side of (3), for every $\bar{c}_2$, attains its maximum at $\bar{y}_2 = \bar{a}_2$, by the triangle inequality. This gives the following algorithm algorithm for our problem: for every $\bar{a} \in S^+$, check if $\bar{y}_2 = \bar{a}_2$ satisfies all inequalities in (2). If for some $\bar{a} \in S^+$ it does, the set X is not a sufficient reason, otherwise X is a sufficient reason.

The argument for the case $f^1_{S^+,S^-}((\bar{x}_1, \bar{x}_2)) = 0$ is exactly the same, with the roles of S^+ and S^- swapped and with non-strict inequalities replaced by strict ones. ◻

As a corollary, we obtain the following:

Corollary 3. *Consider the setting* $(\mathbb{R}, D_1)$*. There is a polynomial time algorithm that, given sets* $S^+, S^- \subseteq \mathbb{R}^n$ *and a vector* $\bar{x} \in \mathbb{R}^n$*, it computes a minimal sufficient reason* $\bar{y}$ *for* $\bar{x}$ *with respect to* $f^1_{S^+,S^-}$*.*

The complexity of k-Check Sufficient Reason$(\mathbb{R}, D_1)$ and k-Minimal Sufficient Reason$(\mathbb{R}, D_1)$ for $k \geq 3$ remains open.

7 Results on the Discrete Setting

We consider the metric space family $(\{0, 1\}, D_H)$, where D_H is the Hamming distance on $\{0, 1\}^n$, for every $n > 0$. Our results for the discrete setting resemble those for the continuous setting under the ℓ_1-distance; however, our proof techniques differ. Additionally, we show that, in the discrete setting, the k-Check Sufficient Reason problem is NP-hard for every odd integer $k \geq 3$. Our hardness result holds even when the input subset of components X is the empty set. This directly implies hardness for k-Minimal Sufficient Reason and for the problem of computing a minimal sufficient reason, when $k \geq 3$.

Counterfactual explanations We prove that k-Counterfactual Explanation is intractable for every odd integer $k \geq 1$, employing different techniques than those used for the NP-hardness of the problem in the continuous setting under the ℓ_1-distance. Unlike the continuous case, where Theorem 3 shows NP-hardness even when $|S^+| = |S^-| = 1$, this does not hold in the discrete setting. We address this by constructing a reduction with an unbounded size for S^+.

Theorem 4. *The problem k-Counterfactual Explanation$(\{0,1\}, D_H)$ is* NP*-complete for every fixed odd integer $k \geq 1$.*

To prove this result, we first establish the intractability of the following intermediate problem. Fix an integer $p \geq 0$. The input to the problem is given by a Boolean matrix B of dimension $m \times n$ and an integer $\ell \leq n$. We want to know if it is possible to find a set $T \subseteq \{1, \dots, n\}$ of size $|T| \leq \ell$ such that, after *flipping* every column from B whose index is in T, the *weight* of at least $n - p$ rows in the resulting matrix is at most $|T| - 1$. Here, "flipping a column" means replacing the 0s with 1s and vice versa in that column's elements. Additionally, the weight of a row is defined as the number of 1s it contains. We call this problem *p-Boolean Matrix Column Flipping*, or simply p-BMCF.

Proposition 5. *p-BMCF is* NP*-complete, for every $p \geq 0$.*

Proof. Fix $p \geq 0$. We reduce from the following modified version of the *Vertex Cover* problem: Given an undirected graph $G = (V, E)$ and an integer ℓ, is there a set $V' \subseteq V$ with $|V'| \leq \ell$ such that at least $|E| - p$ edges have an endpoint in V'? Notice that for $p = 0$ this is exactly the Vertex Cover problem. The fact that this problem is NP-hard for $p > 0$ is obtained by an easy reduction from Vertex Cover (simply extend the input graph with p additional isolated edges).

We define a Boolean matrix A as the transpose of the incidence matrix of G, i.e., $A[i, j] = 1$ if and only if edge e_i is incident to vertex v_j. We then define our matrix B by extending A with a column of all 1s on the right. We consider the input $(B, \ell + 1)$ for the problem p-BMCF.

Suppose first that $T \subseteq \{1, \dots, |V| + 1\}$ is a solution for the input $(B, \ell + 1)$ of p-BMCF. Notice, by construction, that $|T| > 0$ (as we can assume without loss of generality that G contains at least $p + 1$ edges, and thus B contains at least $p + 1$ rows). Then there exists a solution $T' \subseteq \{1, \dots, |V| + 1\}$ for $(B, \ell + 1)$ such that $|T| = |T'|$ and T' contains index $|V| + 1$. In fact, if T does not contain index $|V| + 1$ then we can obtain solution T' by simply removing any index from T and adding index $|V| + 1$. Then $|T'| \leq \ell + 1$ and $|T''| \leq \ell$, for $T'' = T' \setminus \{|V| + 1\}$. We claim that $U = \{v_i \mid i \in T''\}$ satisfies that at least $|E| - p$ edges of G have an endpoint in U. In fact, take an arbitrary edge $e \in E$ such that row e has weight at most $|T'| - 1$ after flipping the columns in T'. We know that there exist at least $|E| - p$ such edges. We claim that, for each such an edge e, it is the case that $U \cap e \neq \emptyset$. Assume, on the contrary. Then the weight of row e after flipping the columns in T' is

$$2 - |U \cap e| + (|T'| - 1) - |U \cap e| = |T'| + 1.$$

This is a contradiction.

Suppose, in turn, that G has a vertex set $U \subseteq V$ with $|U| \leq \ell$ such that at least $|E| - p$ edges have an endpoint in U. Let us define $T = \{i \mid v_i \in U\}$ and $T' = T \cup \{|V| + 1\}$. Then $|T'| = |U| + 1 \leq \ell + 1$. We claim that T' is a solution for the input $(B, \ell + 1)$ of p-BMCF. Take an arbitrary edge $e \in E$ that is covered by U. In a similar fashion as above, the weight of row e after flipping the columns in T' is

$$2 - |U \cap e| + (|T'| - 1) - |U \cap e| = |T'| + 1 - 2|U \cap e|.$$

But $|U \cap e| > 0$ since U covers e, which implies that the weight of row e after flipping the columns in T' is at most $|T'| - 1$. ◻

We are now ready to prove Theorem 4

Proof of Theorem 4. Fix an odd integer $k = 2p + 1$, for $p \geq 0$. The proof is by reduction from p-BMCF. Suppose that the input to p-BMCF is given by a Boolean matrix B of dimension $m \times n$ and an integer $\ell \leq n$. From the proof of Proposition 5, we can assume that B contains no repeated rows, each row of B contains at least two 0s, and $m \geq p + 1$. From B, we define the input $(S^+, S^-, \bar{x}, \ell)$ for k-Counterfactual Explanation as follows:

- Each row b of B defines a tuple in S^+ of dimension $n + p + 1$. This tuple is obtained by extending b with $p + 1$ 0s on the right.
- We define S^- to contain all tuples of the form $\{0\}^{n+j} \times \{1\} \times \{0\}^{p-j}$, for $1 \leq j \leq p + 1$. Observe that there are $p + 1$ such tuples.
- Finally, $\bar{x} = \{1\}^{n+p+1}$.

Notice that the $p + 1$ closest points to $\bar{x}$ in the set $S^+ \cup S^-$ belong to S^+, as every row of B contains at least two 0s. Therefore, $f^k_{S^+,S^-}(\bar{x}) = 1$.

Assume first that the input (B, ℓ) to p-BMCF has a solution $T \subseteq \{1, \ldots, n\}$. Then $|T| \leq \ell$. Let $\bar{y} \in \{0, 1\}^{n+p}$ be the point that is obtained from $\bar{x}$ by flipping from 1 to 0 precisely the elements indexed in T. Then the Hamming distance between $\bar{x}$ and $\bar{y}$ is $|T| \leq \ell$, and the Hamming distance between $\bar{y}$ and any point in S^- is $n - |T| + p$. Let B_T be the matrix that is obtained from B after flipping the columns in T. We know that at least $m - p$ of the rows in B_T have weight at most $|T| - 1$. Take any such row b, and let $\phi(b)$ be its corresponding element in S^+. It follows that the distance between $\bar{y}$ and $\phi(b)$ is at least $n - (|T| - 1) + p + 1 = n - |T| - p + 2 > n - |T| + p$. Therefore, S^+ contains at most p elements which are at distance at most $n - |T| + p$ from $\bar{y}$. Hence, $f^k_{S^+,S^-}(\bar{y}) = 0$.

Assume, in turn, that the input $(S^+, S^-, \bar{x}, \ell)$ to the problem Counterfactual Explanation has a solution given by $\bar{y}$, where $\bar{y}$ is at distance $d \leq \ell$ from $\bar{x}$. Let $T \subseteq \{1, n + p + 1\}$ be the set of indices for which $\bar{y}$ takes value 0 and $T' = T \cap \{1, \ldots, n\}$. We claim that T' is a solution for (B, ℓ) with respect to p-BCMF. Notice first that the Hamming distance between $\bar{y}$ and an arbitrary element in S^- is at most

$$(n - |Y'|) + (p + 1) - (d - |Y'|) + 1 \ = \ n + p - d + 2.$$

This means that there cannot be $p+1$ elements in S^+ whose Hamming distance from $\bar{y}$ is at most $n+p-d+2$ (as, otherwise, $f^k_{S^+,S^-}(\bar{y}) = 1$, which is a contradiction). Suppose, for the sake of contradiction, that after flipping the columns in T' and obtaining matrix $B_{T'}$ there are at least $p + 1$ rows whose weight is at least $|Y'|$. Take any such row b. Then the distance between $\bar{y}$ and the element in S^+ that uniquely represents b is at most

$$(n - |Y'|) + (p + 1) - (d - |Y'|) \ = \ n + p - d + 1 \ < \ n + p - d + 2.$$

This is our desired contradiction. □

Abductive explanations We start by establishing that the problems 1-Check Sufficient Reason and 1-Minimal Sufficient Reason are tractable in the discrete setting.

Proposition 6. *The problem* 1-Check Sufficient Reason$(\{0, 1\}, D_H)$*, and hence also* 1-Minimal Sufficient Reason$(\{0, 1\}, D_H)$*, is polynomial-time solvable.*

Proof. We show that 1-Check Sufficient Reason$(\{0, 1\}, D_H)$ can be solved in polynomial time. Let $\bar{x} \in \{0, 1\}^n$, $S^+, S^- \subseteq \{0, 1\}^n$, and $X \subseteq \{1, \ldots, n\}$. Suppose that X is not a sufficient reason. Assume without loss of generality that $f^1_{S^+,S^-}(\bar{x}) = 1$. Then there is a vector $\bar{z} \in \{0, 1\}^n$ with $\bar{z}[i] = \bar{x}[i]$ for every $i \in X$ such that $f^1_{S^+,S^-}(\bar{z}) = 0$. For each $\bar{y} \in S^-$, let $\bar{y}_X$ be the vector such that $\bar{y}_X[i] = \bar{x}[i]$, for every $i \in X$, and $\bar{y}_X[i] = \bar{y}[i]$, for every $i \notin X$. We claim that without loss of generality, the vector $\bar{z}$ can always be chosen to be a vector in $\{\bar{y}_X \mid \bar{y} \in S^-\}$. Indeed, since $f^1_{S^+,S^-}(\bar{z}) = 0$, there is $\bar{y} \in S^-$ such that $d_H(\bar{z}, \bar{y}) < d_H(\bar{z}, \bar{w})$, for every $\bar{w} \in S^+$. By flipping the components of $\bar{z}$ belonging to $\{i \notin X \mid \bar{z}[i] \neq \bar{y}[i]\}$ to obtain $\bar{y}_X$, the distance $d_H(\bar{z}, \bar{y})$ decreases by an amount of $r = |\{i \notin X \mid \bar{z}[i] \neq \bar{y}[i]\}|$, while the distances $d_H(\bar{z}, \bar{w})$ can decrease by at most a quantity of r. Hence the strict inequality still holds, in particular, we have that $d_H(\bar{y}_X, \bar{y}) < d_H(\bar{y}_X, \bar{w})$, for every $\bar{w} \in S^+$.

The above this discussion implies that in order to check whether X is a sufficient reason when $f^1_{S^+,S^-}(\bar{x}) = 1$ (the other case is analogous), it suffices to check that none of the vectors in $\{\bar{y}_X \mid \bar{y} \in S^-\}$ satisfies $f^1_{S^+,S^-}(\bar{y}_X) = 0$. This can be easily checked in polynomial time. □

As a corollary, we obtain the following:

Corollary 4. *Consider the discrete setting* $(\{0,1\}, D_H)$. *There is a polynomial time algorithm that, given sets* $S^+, S^- \subseteq \mathbb{R}^n$ *and a vector* $\bar{x} \in \mathbb{R}^n$, *it computes a minimal sufficient reason* $\bar{y}$ *for* $\bar{x}$ *with respect to* $f^1_{S^+,S^-}$.

In contrast, both k-Check Sufficient Reason and k-Minimal Sufficient Reason become NP-hard in this setting when $k \geq 3$.

Theorem 5. *The problems* k-Check Sufficient Reason$(\{0,1\}, D_H)$ *and* k-Minimal Sufficient Reason$(\{0,1\}, D_H)$ *are* NP-*hard for every odd integer* $k \geq 3$.

We stress that Theorem 5 also implies hardness for computing minimal reasons, in the case $k \geq 3$. Indeed, the proof of Theorem 5 shows hardness even when the input subset of components X is the empty set. Computing a minimal sufficient reason would imply a solution to that problem: we only need to check whether that the obtained minimal sufficient reason is empty or not.

Proof of Theorem 5. We show that k-Check Sufficient Reason$(\{0,1\}, D_H)$ is NP-hard for every fixed odd integer $k \geq 3$. Fix $c \geq 0$, we reduce from the following variant of the Vertex Cover problem: given a connected undirected graph $G = (V, E)$ with n vertices, for an even integer $n > 2(c+1)$, decide whether there is a subset of vertices $C \subseteq V$ with size $|C| \leq n/2$ such that at least $|E| - c$ edges have an endpoint in C. To see that this problem is NP-hard, recall from the proof of Proposition 5 that the following problem in NP-hard ($c \geq 0$ is fixed): given an undirected graph $G = (V, E)$ and an integer $\ell \geq 0$, decide whether there is a subset of vertices $C \subseteq V$ with size $|C| \leq \ell$ such that at least $|E| - c$ edges have an endpoint in C. Clearly this problem remains hard even when the number of vertices of G is an even integer $n > 2(c+1)$ and $c < \ell < n - c - 1$. We can reduce from this problem to our problem posed above. Consider an instance (G, ℓ), where $G = (V, E)$ and n is the number of vertices. We consider two cases:

1. Suppose that $\ell \geq n/2$. The graph $G' = (V', E')$ is obtained from G by first adding $2\ell - n$ isolated vertices, and then adding two new vertices u, v and new edges such that u is adjacent to all the other vertices in G'. We have $|V'| = 2\ell + 2$. Suppose there is a subset $C \subseteq V$ with $|C| \leq \ell$ such that at least $|E| - c$ edges in G have an endpoint in C. Then the set $C' = C \cup \{u\}$ also covers at least $|E'| - c$ edges in G' and its size is $|C'| = |C| + 1 \leq \ell + 1 = |V'|/2$. On the other hand, suppose there is a subset $C' \subseteq V'$ with $|C'| \leq |V'|/2 = \ell + 1$ such that at least $|E'| - c$ edges in G' have an endpoint in C'. It must hold that $u \in C'$, otherwise, since the number of edges incident to u is $2\ell + 1$ and $|C'| \leq \ell + 1 = 2\ell + 1 - \ell < 2\ell + 1 - c$, there would more than c edges incident to u not being covered by C'. Then $C = C' \cap V$ covers at least $|E| - c$ edges in G and satisfies that $|C| \leq \ell$.

2. Suppose that $\ell < n/2$. In this case, the graph $G' = (V', E')$ is obtained from G by adding a set K of $n - 2\ell$ new vertices and new edges such that every vertex $u \in K$ is adjacent to all the other vertices in G'. We have $|V'| = 2n - 2\ell$. Suppose there is a subset $C \subseteq V$ with $|C| \leq \ell$ such that at least $|E| - c$ edges in G have an endpoint in C. Then the set $C' = C \cup K$ also covers at least $|E'| - c$ edges in G' and its size is $|C'| = |C| + n - 2\ell \leq n - \ell = |V'|/2$. On the other hand, suppose there is a subset $C' \subseteq V'$ with $|C'| \leq |V'|/2 = n - \ell$ such that at least $|E'| - c$ edges in G' have an endpoint in C'. It must hold that $K \subseteq C'$, otherwise, if $u \in K \setminus C'$, since the number of edges incident to u is $2n - 2\ell - 1$, then the number of edges not being covered by C' would be at least $2n - 2\ell - 1 - |C'| \geq n - \ell - 1 > c$. Then $C = C' \cap V$ covers at least $|E| - c$ edges in G and satisfies that $|C| \leq n - \ell - |K| \leq \ell$.

Fix an odd integer $k \geq 3$. We present a reduction from the variant of Vertex Cover posed above, with $c = (k-1)/2$, to k-Check Sufficient Reason$(\{0,1\}, D_H)$. Let $G = (V, E)$ be a connected undirected graph with n vertices, for an even integer $n > 2(c+1)$. Assume that $V = \{1, \ldots, n\}$ and $E = \{e_1, \ldots, e_m\}$, for $m \geq 1$. The vector dimension is $n + (k-1)/2$ and then we write vectors in $\{0,1\}^{n+(k-1)/2}$ as concatenations $\bar{x}\bar{y}$ of two vectors $\bar{x} \in \{0,1\}^n$ and $\bar{y} \in \{0,1\}^{(k-1)/2}$. We denote by $\bar{0}_r$ and $\bar{1}_r$ the vectors $(0, \ldots, 0) \in \{0,1\}^r$ and $(1, \ldots, 1) \in \{0,1\}^r$, respectively. For $h \in \{1, \ldots, (k-1)/2\}$ we define the canonical vector $\bar{\alpha}_h \in \{0,1\}^{(k-1)/2}$ such that $\bar{\alpha}_h[h'] = 1$ for $h' = h$ and $\bar{\alpha}_h[h'] = 0$, for $h' \neq h$. For each $j \in \{1, \ldots, m\}$, we define the vector $\bar{y}_j \in \{0,1\}^n$ such that $\bar{y}_j[i] = 1$ if the edge e_j is incident to the vertex i, and $\bar{y}_j[i] = 0$ otherwise. The sets

$S^-, S^+ \subseteq \{0,1\}^{n+(k-1)/2}$ are defined as follows:

$$S^- = \{\bar{y}_j\bar{\alpha}_1 \mid j \in \{1,\dots,m\}\}$$
$$S^+ = \{\bar{0}_n\bar{\alpha}_1\} \cup \{\bar{1}_n\bar{\alpha}_h \mid h \in \{1,\dots,(k-1)/2\}\}.$$

Finally, we set $\bar{x} = \bar{0}_n\bar{0}_{(k-1)/2}$. Note that $f^k_{S^+,S^-}(\bar{x}) = 0$ as $d_H(\bar{x},\bar{y}) = 3$ for every $\bar{y} \in S^-$, while $d_H(\bar{x},\bar{0}_n\bar{\alpha}_1) = 1$ and $d_H(\bar{x},\bar{1}_n\bar{\alpha}_h) = n+1$. We claim that there is a subset of vertices $C \subseteq \{1,\dots,n\}$ with size $|C| \leq n/2$ such that at least $|E| - c$ edges have an endpoint in C if and only if the empty set is not a sufficient reason for $\bar{x}$ w.r.t $f^k_{S^+,S^-}$. The latter condition is equivalent to the existence of a vector $\bar{z} \in \{0,1\}^{n+(k-1)/2}$ with $f^k_{S^+,S^-}(\bar{z}) = 1$.

Suppose first that there is such a set $C \subseteq \{1,\dots,n\}$ and assume without loss of generality that $|C| = n/2$. Define $\bar{w}_C \in \{0,1\}^n$ such that $\bar{w}_C[i] = 0$ if $i \in C$ and $\bar{w}_C[i] = 1$ otherwise. We claim that $f^k_{S^+,S^-}(\bar{z}) = 1$ for the vector $\bar{z} = \bar{w}_C\bar{0}_{(k-1)/2}$. Using Proposition 1, it suffices to provide a set $A \subseteq S^+$ of $(k+1)/2$ vectors and a set $B \subseteq S^-$ of at most $(k-1)/2$ vectors such that $d_H(\bar{z},\bar{a}) \leq d_H(\bar{z},\bar{b})$, for every $\bar{a} \in A$ and $\bar{b} \in S^- \setminus B$. We set $A = S^+$ and B as:

$$B = \{\bar{y}_j\bar{\alpha}_1 \mid e_j \text{ is not covered by } C\}.$$

Note that $|A| = (k+1)/2$ and $|B| \leq c = (k-1)/2$. Since $d_H(\bar{w}_C,\bar{0}_n) = d_H(\bar{w}_C,\bar{1}_n) = n/2$ and $d_H(\bar{0}_{(k-1)/2},\bar{\alpha}_h) = 1$, it follows that $d_H(\bar{z},\bar{a}) = 1 + n/2$ for all $\bar{a} \in A$. On the other hand, consider $\bar{b} = \bar{y}_j\bar{\alpha}_1 \in S^- \setminus B$. By definition e_j is covered by C. It follows that:

$$|\{i \in \{1,\dots,n\} \mid i \in C, i \in e_j\}| \geq 1 \qquad |\{i \in \{1,\dots,n\} \mid i \notin C, i \in e_j\}| \leq 1$$

Then:

$$\begin{aligned} d_H(\bar{w}_C,\bar{y}_j) &= |\{i \mid \bar{w}_C[i] = 0, \bar{y}_j[i] = 1\}| + |\{i \mid \bar{w}_C[i] = 1, \bar{y}_j[i] = 0\}| \\ &= |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C, i \notin e_j\}| \\ &= |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C\}| - |\{i \mid i \notin C, i \in e_j\}| \\ &\geq 1 + n/2 - 1 \end{aligned}$$

We conclude that $d_H(\bar{z},\bar{b}) = 1 + d_H(\bar{w}_C,\bar{y}_j) \geq 1 + n/2$, and then $f^k_{S^+,S^-}(\bar{z}) = 1$.

Now suppose that there exists a vector $\bar{z} = \bar{w}\bar{\alpha}$ such that $f^k_{S^+,S^-}(\bar{z}) = 1$. Again using Proposition 1, it follows that there is a set $B \subseteq S^-$ such that $d_H(\bar{z},\bar{a}) \leq d_H(\bar{z},\bar{b})$, for every $\bar{a} \in S^+$ and $\bar{b} \in S^- \setminus B$. Define $C = \{i \in \{1,\dots,n\} \mid \bar{w}[i] = 0\}$. Let $r = d_H(\bar{\alpha},\bar{\alpha}_1)$. We prove that C covers each edge e_j with $\bar{y}_j\bar{\alpha}_1 \in S^- \setminus B$, and then C covers at least $m - (k-1)/2 = m - c$ edges. By contradiction, assume that C fails to cover e_j, where $\bar{y}_j\bar{\alpha}_1 \in S^- \setminus B$. We claim that $d_H(\bar{z},\bar{0}_n\bar{\alpha}_1) > d_H(\bar{z},\bar{y}_j\bar{\alpha}_1)$, which is a contradiction as $\bar{0}_n\bar{\alpha}_1 \in S^+$. We have that

$$d_H(\bar{z},\bar{0}_n\bar{\alpha}_1) = r + d_H(\bar{w},\bar{0}_n) = r + |\{i \mid \bar{w}[i] = 1\}| = r + |\{i \mid i \notin C\}|.$$

On the other hand, we have that

$$\begin{aligned} d_H(\bar{z},\bar{y}_j\bar{\alpha}_1) &= r + d_H(\bar{w},\bar{y}_j) \\ &= r + |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C, i \notin e_j\}| \\ &= r + |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C\}| - |\{i \mid i \notin C, i \in e_j\}| \\ &= r + 0 + |\{i \mid i \notin C\}| - 2 \\ &< r + |\{i \mid i \notin C\}|. \end{aligned}$$

Then if $|C| \leq n/2$, we are done. We show that $|C| \geq n/2 + 1$ cannot happen. Suppose first that $|C| \geq n/2 + 2$. We claim that $d_H(\bar{z},\bar{1}_n\bar{\alpha}_1) > d_H(\bar{z},\bar{y}_j\bar{\alpha}_1)$, for $\bar{1}_n\bar{\alpha}_1 \in S^+$ and every $y_j\bar{\alpha}_1 \in S^-$. Indeed, it holds that:

$$d_H(\bar{z},\bar{1}_n\bar{\alpha}_1) = r + d_H(\bar{w},\bar{1}_n) = r + |\{i \mid \bar{w}[i] = 0\}| = r + |C| \geq r + n/2 + 2.$$

On the other hand, we have:

$$\begin{aligned}
d_H(\bar{z}, \bar{y}_j\bar{\alpha}_1) &= r + d_H(\bar{w}, \bar{y}_j)\\
&= r + |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C, i \notin e_j\}|\\
&= r + |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C\}| - |\{i \mid i \notin C, i \in e_j\}|\\
&\leq r + 2 + |\{i \mid i \notin C\}|\\
&\leq r + 2 + n/2 - 2\\
&< r + n/2 + 2.
\end{aligned}$$

Finally, suppose $|C| = n/2 + 1$. By assumption in our variant of Vertex Cover, we have that G is connected and $n > 2(c+1)$. This implies that there are more than c edges with an endpoint not in C. Indeed, the number of vertices in $V \setminus C$ is $n/2 - 1 > c$. Suppose the subgraph of G induced by $V \setminus C$ has s connected components. By connectivity of G, there must be an edge between each of these connected components and C. Hence the number of edges with an endpoint in $V \setminus C$ is at least $|V \setminus C| - s + s > c$. It follows that there exists a vector $\bar{y}_j\bar{\alpha}_1 \in S^- \setminus B$ such that e_j has an endpoint not in C. We have that $d_H(\bar{z}, \bar{1}_n\bar{\alpha}_1) > d_H(\bar{z}, \bar{y}_j\bar{\alpha}_1)$, which is a contradiction as $\bar{1}_n\bar{\alpha}_1 \in S^+$. Indeed, we have that:

$$d_H(\bar{z}, \bar{1}_n\bar{\alpha}_1) = r + |C| = r + n/2 + 1.$$

On the other hand, we have that:

$$\begin{aligned}
d_H(\bar{z}, \bar{y}_j\bar{\alpha}_1) &= r + |\{i \mid i \in C, i \in e_j\}| + |\{i \mid i \notin C\}| - |\{i \mid i \notin C, i \in e_j\}|\\
&\leq r + 1 + |\{i \mid i \notin C\}|\\
&= r + 1 + n/2 - 1\\
&< r + n/2 + 1.
\end{aligned}$$

Hardness for k-Minimal Sufficient Reason($\{0,1\}, D_H$), when $k \geq 3$, follows directly from the same reduction. This concludes the proof of the theorem. □

8 Implementation and Experiments

The results presented in the previous sections offer a comprehensive understanding of the requirements for computing explanations for k-NN classifiers in practice. In this section, we explore this topic further through a preliminary practical analysis, focusing on the widely used case of $k = 1$. This case is particularly appealing as it reduces implementation complexity even for problems that remain tractable for larger values of k. We examine the problems of 1-Counterfactual Explanation and 1-Minimal Sufficient Reason.

8.1 Experimental setup and datasets

Our implementation is written in Python (3.10), but naturally calls external solvers. Namely, we use the standard `cvxpy` library for convex programming, the `Gurobi` [27] solver for IQP, and the recent `cardinality-cadical` solver [50] for SAT-solving. All experiments were run on a Macbook Pro M1 2020 with 16GB of RAM. We experiment both on the MNIST dataset of handwritten digit recognition [19] and on synthetic random data. For the MNIST dataset, we consider both the original grayscale 28×28 images, as well as a binarized version to represent the discrete setting, and different rescalings of the images to experiment with a different number of dimensions. Similarly, we consider subsets of the training data of different sizes (MNIST was originally split into 60 000 training examples and 10 000 test images). When computing an explanation for an image classified as digit $d \in \{0, 1, \ldots, 10\}$, we consider all images of digit d as "positive" examples, and images of digits $d' \neq d$ as "negative". For the synthetic random data, we consider uniformly random vectors in $\{0,1\}^n$, labeled according to independent Bernoulli variables of parameter $p = \frac{1}{2}$, since additional experiments with other values of p displayed a similar behavior.

8.2 Implementation

For computing minimal sufficient reasons over $(\mathbb{R}, D_1)$, we directly implement the simple algorithm from Proposition 4, using the efficient `FAISS` library for fast 1-NN search [20]. For computing counterfactual explanations in $(\mathbb{R}, D_2)$, we implement the convex program from Theorem 2, ignoring tie-breaking concerns for simplicity. Computing counterfactual explanations in $(\mathbb{R}, D_1)$, we defer to the optimized implementation of a mixed integer program by Contardo et al. [16]. For counterfactual explanations in the discrete setting, we consider first the following IQP formulation for a vector $\bar{x}$ classified positively:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} (\bar{x}[i] - \bar{y}[i])^2 \\
\text{subject to } \; & d^+ = \min_{\bar{z} \in S^+} \sum_{i=1}^{n} (\bar{z}[i] - \bar{y}[i])^2, \\
& d^- = \min_{\bar{z} \in S^-} \sum_{i=1}^{n} (\bar{z}[i] - \bar{y}[i])^2, \\
& d^- < d^+, \\
& \bar{y} \in \{0,1\}^d, \; d^+ \in \mathbb{R}, \; d^- \in \mathbb{R},
\end{aligned}
$$

where constraints $m = \min(r_1, \ldots, r_t)$ can be expressed by introducing indicator variables $v_1, \ldots, v_t \in \{0, 1\}$ and adding constraints $m \leq r_i$ and $v_i \cdot r_i \leq m$, for every $i \in \{1, \ldots, t\}$, and $\sum_{i=1}^{t} v_i = 1$.

SAT encoding We also propose a CNF encoding to find a closest counterfactual $\bar{y}$, leveraging the native support for (guarded) cardinality constraints of the recent `cardinality-cadical` solver [50]. Given boolean literals $\ell_1, \ldots, \ell_n$, cardinality constraints are of the form $\sum_{i=1}^{n} \ell_i \geq b$, for some constant integer b (the "bound"). On the other hand, "guarded" cardinality constraints, provided a different boolean literal g, are of the form $g \implies \left(\sum_{i=1}^{n} \ell_i \geq b\right)$. Our encoding uses boolean variables $y_1, \ldots, y_n$ where y_i corresponds to whether $\bar{y}[i] = 1$, and variables $c_1, \ldots, c_{|S^-|}$, where c_i intuitively represents that the i-th point in S^- (under some fixed ordering) will be the closest point to $\bar{y}$ among $S^+ \cup S^-$. We thus add first a clause of the form $\bigvee_{i=1}^{|S^-|} c_i$. Then, if we call $\bar{o}$ to the i-th point in S^-, we need to enforce that

$$
c_i \implies d_H(\bar{y}, \bar{o}) < d_H(\bar{y}, \bar{s}), \quad \forall \bar{s} \in S^+, \tag{4}
$$

which we show next how to encode as a guarded cardinality constraint. Let us focus on a fixed pair $\bar{o}, \bar{s}$, and define the sets

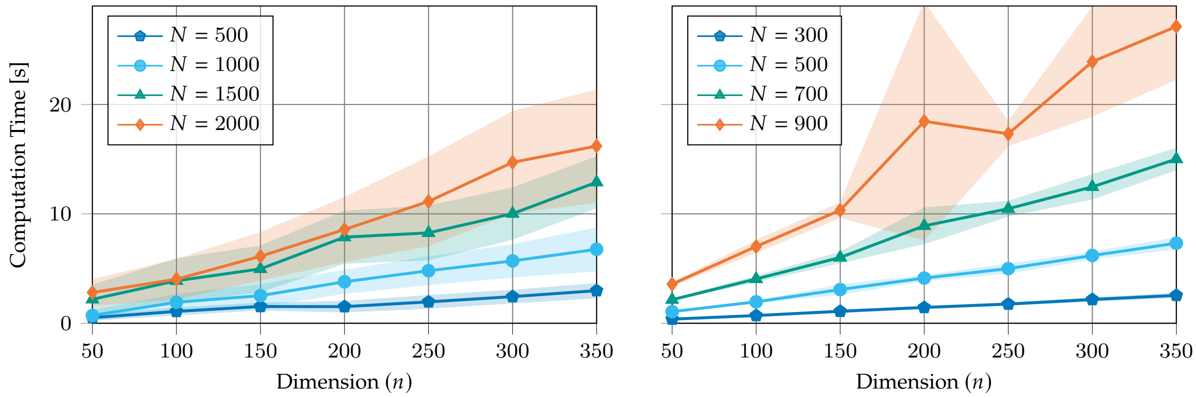
$$
\Delta_0 := \{i \mid \bar{o}[i] = 0, \bar{s}[i] = 1\} \quad ; \quad \Delta_1 := \{i \mid \bar{o}[i] = 1, \bar{s}[i] = 0\}.
$$

We then have the following equivalence:

$$
\begin{aligned}
d_H(\bar{y}, \bar{o}) < d_H(\bar{y}, \bar{s}) &\iff \sum_{i \in \Delta_0} y_i + \sum_{i \in \Delta_1} \neg y_i < \sum_{i \in \Delta_0} \neg y_i + \sum_{i \in \Delta_1} y_i \\
&\iff \sum_{i \in \Delta_0} y_i + \sum_{i \in \Delta_1} (1 - y_i) < \sum_{i \in \Delta_0} (1 - y_i) + \sum_{i \in \Delta_1} y_i \\
&\iff |\Delta_1| - |\Delta_0| < 2 \sum_{i \in \Delta_1} y_i - 2 \sum_{i \in \Delta_0} y_i \\
&\iff \frac{|\Delta_0 + \Delta_1|}{2} < \sum_{i \in \Delta_0} \neg y_i + \sum_{i \in \Delta_1} y_i,
\end{aligned}
$$

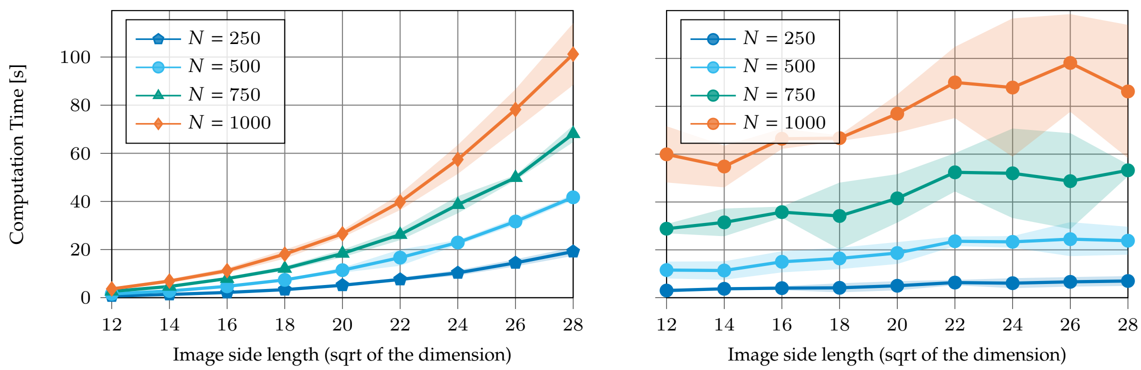
which implies that we can encode Equation (4) as the following guarded cardinality constraint:

$$
c_i \implies \left(\sum_{i \in \Delta_0} \neg y_i + \sum_{i \in \Delta_1} y_i \right) \geq \left\lfloor \frac{|\Delta_0 + \Delta_1|}{2} \right\rfloor + 1.
$$



(a) IQP (b) SAT

Figure 3: Runtimes for counterfactual explanations over $\{0,1\}^n$. The total training set has size $N := |S^+|+|S^-|$, consisting of independent uniformly random samples from $\{0,1\}^n$. Confidence intervals of 95% over 30 independent runs are displayed.



(a) Minimal Sufficient Reason (ℓ_1) (b) Counterfactual (ℓ_2)

Figure 4: Runtimes for explanations over the MNIST dataset. The training set used has size $N := |S^+|+|S^-|$. Confidence intervals of 95% over 5 independent runs are displayed.

The encoding therefore has $|S^+| \cdot |S^-|$ guarded cardinality constraints. Finally, in order to minimize the distance $d_H(\bar{x}, \bar{y})$, note that a cardinality constraint

$$\sum_{i \text{ s.t. } \bar{x}[i]=0} y_i + \sum_{i \text{ s.t. } \bar{x}[i]=1} \neg y_i \geq n - k$$

is equivalent to $d_H(\bar{x}, \bar{y}) \leq k$. Therefore, by doing a binary search over the parameter k (or a linear search if the answer is expected to be small) we obtain a closest counterfactual explanation.

Results. On the discrete setting, Figure 3 displays experimental results over synthetic random data, comparing the performance of IQP and SAT solving. While our implementation of the former scales significantly better than the latter, it is worth mentioning that `Gurobi` was run using 8 threads, whereas `cardinality-cadical` is a single-threaded program. On the continuous setting, Figure 4 displays experimental results for the algorithms described in Theorem 2 and Proposition 4. We remark that the use of a library for fast NN-classification such as FAISS [20] was key for performance in the computation of minimal sufficient reasons. In general, our results suggest that for hundreds of features, and up to a thousand points, explanations can be computed in practice. A faster implementation in a lower level language, using pruning heuristics as those of [16,21], is part of our future work.

9 Final Remarks

Our work represents an initial step in studying the computational cost of generating explanations for k-NN classifiers. As demonstrated, the landscape is nuanced, with the complexity of finding explanations varying depending on the metric used. We believe that our results and proof techniques provide valuable insights into the types of methods needed to practically address the explanation problems studied in this paper. A summary of our results is shown in Table 1.

The kind of explanations we have studied are often said to be "local" [26,28], since they aim to elucidate the behavior of a classifier in a local region of the space around an input point, as opposed to "global" explainability which aims to provide insight into a classier as a whole. Recently, Bassan et al. studied the difference between local and global interpretability from a computational-complexity perspective [12]. Interestingly, a recent line of work has studied the computational problem of *thinning* k-NN classifiers by removing redundant points in the training set [21,23,52]. Arguably, this line of work contributes to the global interpretability of k-NN classifiers, and in practice might serve to speed up the computation of local explanations.

There are several intriguing directions for future research. First, we aim to clarify the complexity of k-Check Sufficient Reason for $k \geq 3$ in the continuous setting under the ℓ_1-distance. Second, we seek to explore k-Counterfactual Explanation for metrics based on ℓ_p, where $p > 2$. Specifically, we ask whether ℓ_2 is the only metric for which this problem is tractable. Lastly, we are interested in determining the extent to which the NP-hard problems discussed in this paper can be approximated. For instance, can k-Minimum Sufficient Reason, which is NP-hard in all the settings considered, be tackled using polynomial-time approximation algorithms that produce a sufficient reason whose size is reasonably close to the minimum?

Table 1: Summary of complexity results.

Explanation	Counterfactual	Minimal Sufficient Reason		Minimum Sufficient Reason
Metric space	$k \geq 1$	$k = 1$	$k > 1$	$k \geq 1$
$(\mathbb{R}, D_2)$	P (Thm. 2)	P (Thm. 3)	P (Thm. 3)	NP-c (Thm. 1)
$(\mathbb{R}, D_1)$	NP-c (Thm. 3)	P (Prop. 4)	Open	NP-h (Thm. 1)
$(\{0,1\}, D_H)$	NP-c (Thm. 4)	P (Prop. 6)	NP-h (Thm. 5)	NP-h (Thm. 1)

References

[1] Pankaj K. Agarwal, Boris Aronov, Sariel Har-Peled, Jeff M. Phillips, Ke Yi, and Wuzhou Zhang. Nearest neighbor searching under uncertainty II. In *PODS*, pages 115–126. ACM, 2013.

[2] Pankaj K. Agarwal, Alon Efrat, Swaminathan Sankararaman, and Wuzhou Zhang. Nearest-neighbor searching under uncertainty. In *PODS*, pages 225–236. ACM, 2012.

[3] Pankaj K. Agarwal, Kyle Fox, Kamesh Munagala, and Abhinandan Nath. Parallel algorithms for constructing range and nearest-neighbor searching data structures. In *PODS*, pages 429–440. ACM, 2016.

[4] Marcelo Arenas, Daniel Báez, Pablo Barceló, Jorge Pérez, and Bernardo Subercaseaux. Foundations of Symbolic Languages for Model Interpretability. In *Advances in Neural Information Processing Systems*, volume 34, pages 11690–11701. Curran Associates, Inc., 2021.

[5] Marcelo Arenas, Pablo Barceló, Miguel A. Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees. In *NeurIPS*, 2022.

[6] Marcelo Arenas, Pablo Barceló, Diego Bustamante, Jose Caraball, and Bernardo Subercaseaux. A Uniform Language to Explain Decision Trees. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 60–70, 8 2024.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[8] André Artelt and Barbara Hammer. On the computation of counterfactual explanations–a survey. *arXiv preprint arXiv:1911.07749*, 2019.

[9] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of Boolean decision trees. *Data & Knowledge Engineering*, 142:102088, November 2022.

[10] Gilles Audemard, Jean-Marie Lagniez, and Pierre Marquis. On the computation of contrastive explanations for boosted regression trees. In *ECAI*, pages 1083–1091, 2024.

[11] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *NeurIPS*, pages 15487–15498, 2020.

[12] Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability: A computational complexity perspective, 2024.

[13] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is ”nearest neighbor” meaningful? In *ICDT*, pages 217–235, 1999.

[14] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

[15] Guy Blanc, Jane Lange, and Li-Yang Tan. Provably efficient, succinct, and precise explanations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021.

[16] Claudio Contardo, Ricardo Fukasawa, Louis-Martin Rousseau, and Thibaut Vidal. Optimal counterfactual explanations for k-nearest neighbors using mathematical optimization and constraint programming. In *ISCO*, pages 318–331, 2024.

[17] Adnan Darwiche. Logic for explainable AI. In *LICS*, pages 1–11. IEEE, 2023.

[18] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. *CoRR*, abs/2002.09284, 2020.

[19] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[20] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[21] David Eppstein. Finding Relevant Points for Nearest-Neighbor Classification. In *2022 Symposium on Simplicity in Algorithms (SOSA)*, Proceedings, pages 68–78. Society for Industrial and Applied Mathematics, January 2022.

[22] Austen Z. Fan and Paraschos Koutris. Certifiable robustness for nearest neighbor classifiers. In *ICDT*, LIPIcs, pages 6:1–6:20, 2022.

[23] Alejandro Flores-Velazco. Improved search of relevant points for nearest-neighbor classification. *arXiv preprint arXiv:2203.03567*, 2022.

[24] Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Explainable data-driven optimization: From context to decision and back again. In *ICML*, volume 202, pages 10170–10187. PMLR, 2023.

[25] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 36(5):1770–182, 2022.

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018.

[27] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.

[28] Hajar Hakkoum, Ali Idri, and Ibtissam Abnane. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data. *Engineering Applications of Artificial Intelligence*, 131:107829, May 2024.

[29] Yunqiu Han, Yizuo Chen, and Adnan Darwiche. On the complexity of counterfactual reasoning. In *IJCAI*, pages 5676–5684. ijcai.org, 2023.

[30] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. What is the nearest neighbor in high dimensional spaces? In *VLDB*, pages 506–515, 2000.

[31] Jean-Baptiste Hiriart-Urruty and Claude LemarÃ©chal. *Fundamentals of Convex Analysis*. Springer, Berlin, Heidelberg, 2001.

[32] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and João Marques-Silva. On efficiently explaining graph-based classifiers. In *KR*, pages 356–367, 2021.

[33] Alexey Ignatiev and Joao Marques-Silva. SAT-based rigorous explanations for decision lists. In *Theory and Applications of Satisfiability Testing – SAT 2021*, pages 251–269. Springer International Publishing, 2021.

[34] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. From contrastive to abductive explanations and back again. In *AIxIA 2020 – Advances in Artificial Intelligence*, pages 335–355. Springer International Publishing, 2021.

[35] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519, 2019.

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

[37] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *Journal of Artificial Intelligence Research*, 75:261–321, September 2022.

[38] Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and J. Marques-Silva. Efficient explanations with relevant sets. *ArXiv*, abs/2106.00546, 2021.

[39] Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. Provably precise, succinct and efficient explanations for decision trees, 2022.

[40] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. When large language models meet vector databases: A survey. *CoRR*, abs/2402.01763, 2024.

[41] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20(5):223–228, January 1980.

[42] François Lamothe, Claudio Contardo, and Matthieu Gruson. Tilted inequalities and facets of the set covering polytope: A theoretical analysis. *Discret. Appl. Math.*, 357:143–160, 2024.

[43] Harry R Lewis and Christos H Papadimitriou. Elements of the theory of computation, 1997.

[44] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[45] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *NeurIPS*, 2020.

[46] Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *ICML*, pages 7469–7479, 2021.

[47] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *ICML*, pages 7469–7479, 2021.

[48] Joao Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through Formal XAI. In *AAAI*, 2022.

[49] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[50] Joseph E Reeves, Marijn JH Heule, and Randal E Bryant. From clauses to klauses. In *International Conference on Computer Aided Verification*, pages 110–132. Springer, 2024.

[51] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535, 2018.

[52] Josiah Rohrer and Simon Weber. Reducing nearest neighbor training sets optimally and exactly. *arXiv preprint arXiv:2302.02132*, 2023.

[53] Nick Roussopoulos, Stephen Kelley, and Frédéic Vincent. Nearest neighbor queries. In *SIGMOD*, pages 71–79. ACM Press, 1995.

[54] Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Chichester, June 1998.

[55] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5103–5111. AAAI Press, 2018.

[56] Yufei Tao, Ke Yi, Cheng Sheng, and Panos Kalnis. Quality and efficiency in high dimensional nearest neighbor search. In *SIGMOD*, pages 563–576. ACM, 2009.

[57] Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res.*, 70:351–387, 2021.