# 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

January 22, 2026

## 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 $\bar{v}_1, \ldots, \bar{v}_k$ 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 [Biau and Devroye(2015)]. 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 [Beyer et al.(1999), Hinneburg et al.(2000), Tao et al.(2009)] and its properties when dealing with uncertain data [Agarwal et al.(2012), Agarwal et al.(2013), Fan and Koutris(2022)]. Considerable effort has also been directed toward the development of efficient algorithms and data structures to enable scalable NN queries [Roussopoulos et al.(1995), Agarwal et al.(2016)]. As of late, $k$-NN has also become key to several search and retrieval problems in vector databases [Jing et al.(2024)]. 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 [Arrieta et al.(2020)]. 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 [Marques-Silva and Ignatiev(2022)]. 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 $\bar{x}$ [Ribeiro et al.(2018),Darwiche and Hirth(2020),Ignatiev et al.(2019)]. 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* [Shih et al.(2018)]. 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 [Artelt and Hammer(2019),Guidotti(2022)]. More formally, a *contrastive explanation at distance p from $\bar{x}$*, with respect to a given classifier, is another input $\bar{y}$ such that $\|\bar{x} - \bar{y}\| \leqslant p$ 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 [Molnar(2022)]. However, this view is overly simplistic, as interpretability depends on whether individual instances and their features are understandable [Lipton(2018),Molnar(2022)]. 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 [Izza et al.(2023)]. 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 [Shih et al.(2018), Marques-Silva et al.(2020), Barceló et al.(2020),Izza et al.(2021),Huang et al.(2021),Marques-Silva et al.(2021),Blanc et al.(2021),Marques-Silva et al.(2021),Wäldchen et al.(2021),Arenas et al.(2022),Ignatiev et al.(2021),Ignatiev and Marques-Silva(2021), Izza et al.(2022a), Audemard et al.(2022), Izza et al.(2022b), Han et al.(2023), Darwiche(2023), Audemard et al.(2024)]. 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 [Forel et al.(2023),Lamothe et al.(2024), Magagnini et al.(2025)]. 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.
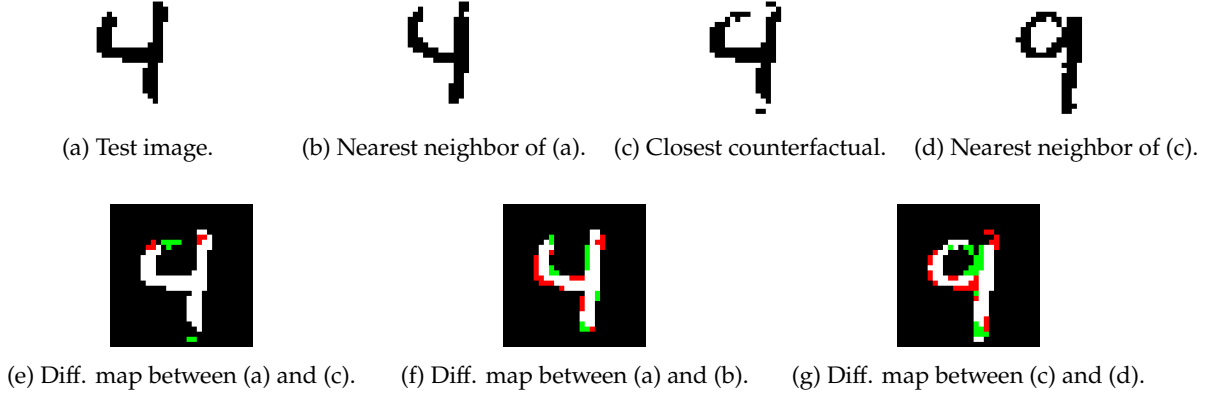
(a) Test image.    (b) Nearest neighbor of (a).    (c) Closest counterfactual.    (d) Nearest neighbor of (c).



(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 $9$.

**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 $\ell_p$-norm for some integer $p > 0$, and (b) the *discrete* setting, where points are Boolean vectors, and the Hamming distance is used.

- 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 terms of the number of features, in both the discrete and continuous settings. In the latter case, hardness holds for every distance $\ell_p$, where $p$ is a positive integer.

- We next examine the continuous setting to assess tractability for other standard explainability problems. Focusing on the $\ell_2$ and $\ell_1$ distances, two of the most commonly used metrics in practice, we show that tractability depends crucially on the metric chosen.

  - For the $\ell_2$-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 of 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. The previous problems can be solved in time $n^{O(k)}$, which is polynomial when $k$ is fixed. For the case of counterfactual explanations, we further show that, under widely believed assumptions from parameterized complexity, the exponential dependence in $k$ is unavoidable, as the problem is W[1]-hard when $k$ is the parameter.

  - We then examine the $\ell_1$-distance and show that not all positive results derived for the $\ell_2$-distance carry over. Specifically, checking the existence of a counterfactual explanation at a certain distance becomes NP-complete 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. For fixed $k \geqslant 3$, we show that checking whether a subset of components of an input $\bar{x}$ is a sufficient reason is coNP-complete, and, as a consequence, that computing minimal sufficient reasons in this case is intractable.

- Afterwards, we consider the discrete setting and show that it resembles the behavior observed for the $\ell_1$-distance in the continuous setting. Specifically: (a) checking the existence of a counterfactual explanation at a certain distance is NP-complete, 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, checking if a subset of components of an input $\bar{x}$ is a sufficient reason for a $k$-NN classifier, for fixed $k \geqslant 3$, is coNP-complete. Our results also imply hardness for computing minimal sufficient reasons, in the case $k \geqslant 3$. We further show that the problem of checking the existence of a sufficient reason of a certain size is complete for the complexity class $\Sigma_2^p$.

- 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 [Contardo et al.(2024)], our SAT encoding is novel and utilizes a recent solver with native support for cardinality constraints [Reeves et al.(2024)]. 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 $\ell_2$-distance, are covered in Section 5, while those for the $\ell_1$-distance are detailed in Section 6. Results for the discrete setting appear in Section 7. Further complexity completeness results are presented in Section 8. Experimental results are discussed in Section 9, while final remarks are included in Section 10.

## 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 $i \in \{1, \ldots, n\}$, we write $\bar{x}[i]$ to denote the $i$th component of vector $\bar{x}$. When it is clear from the context, we also use $\bar{x}_i$ to denote the $i$th 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 $\ell_p$-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 $i \in \{1, \ldots, n\}$ 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*. We assume that $|S^+ \cup S^-| \geqslant k$. For the pair $(M, D)$, we aim to construct a *k-Nearest Neighbor (k-NN) classification function*

$$f_{S^+, S^-}^k : M^n \to \{0, 1\},$$

such that $f_{S^+, S^-}^k(\bar{x}) = 1$ 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 $f_{S^+, S^-}^k(\bar{x}) = 1$ if and only if there is a subset $T \subseteq S^+ \cup S^-$ of size $k$ such

4

that the majority of points of $T$ belong to $S^+$ and $d_n(\bar{x}, \bar{y}) \leqslant d_n(\bar{x}, \bar{z})$ 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 [Contardo et al.(2024)].[1]

In some proofs we use the following characterization of the $k$-NN classification function:

**Proposition 1.** *(a) We have $f_{S^+,S^-}^k(\bar{x}) = 1$ 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 $d_n(\bar{x}, \bar{a}) \leqslant d_n(\bar{x}, \bar{c})$ for every $\bar{a} \in A$ and $\bar{c} \in S^- \setminus B$.*

*(b) We have $f_{S^+,S^-}^k(\bar{x}) = 0$ 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 $\bar{c} \in S^+ \setminus B$.*

*Proof.* Imagine that we start inflating a ball with the center at $\bar{x}$. As the radius grows, some points of $S^+$ and $S^-$ fall inside the ball. We are waiting until either $(k+1)/2$ points of $S^+$ or of $S^-$ fall inside. If this first happens for $S^+$, we set $f_{S^+,S^-}^k(\bar{x}) = 1$, and we set $f_{S^+,S^-}^k(\bar{x}) = 0$ if it first happens for $S^-$. If this happens simultaneously, we set $f_{S^+,S^-}^k(\bar{x}) = 1$, favoring positive classification. In other words, we classify $\bar{x}$ positively if there is a ball that has at least $(k+1)/2$ points of $S^+$ but its interior has at most $(k-1)/2$ points of $S^-$. Likewise, $\bar{x}$ is classified negatively if there is a ball with at least $(k+1)/2$ points of $S^-$ but with at most $(k-1)/2$ points of $S^+$ (in the interior and the boundary). □

## 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 $f_{S^+,S^-}^k$ 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 $f_{S^+,S^-}^k$. We formalize these ideas next using the well-known notion of *sufficient reason*.

Fix an odd integer $k \geqslant 1$. 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 $f_{S^+,S^-}^k$*, if

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

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 $f_{S^+,S^-}^k$ |

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.

---

[1] We focus on the case where $k$ is odd to ensure that ties occur only when the set of $k$ closest points is not unique. If $k$ is even, ties can encompass the entire space (or a significant subset), causing the optimistic tie-breaking rule to classify all points positively, including those in $S^-$. For example, with $k = 2$, if $M^n$ is the real line, $S^+$ the odd numbers, and $S^-$ the even numbers, ties arise everywhere.

| | |
|---|---|
| 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 $|X| \leqslant \ell$ |

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 $k = 1$. 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.* $\square$

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. $\square$

**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 $k \geqslant 1$. 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.

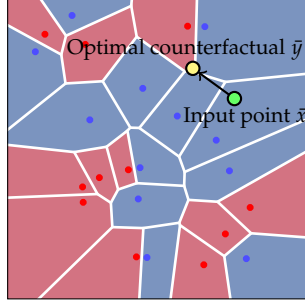| | |
|---|---|
| PROBLEM : | $k$-Counterfactual Explanation$(M, D)$ |
| INPUT : | Two sets $S^+, S^- \subseteq M^n$, a vector $\bar{x} \in M^n$, a rational $\ell > 0$ |
| OUTPUT : | Yes, if there is a counterfactual explanation $\bar{y}$ for $\bar{x}$ with respect to $f^k_{S^+, S^-}$ such that $d_n(\bar{x},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

Figure 2: Illustration of minimum distance counterfactual explanations over $\mathbb{R}^2$ in the $\ell_2$ metric, with $k = 1$. Blue (red) areas are classified negatively (positively).

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 $f_{S^+,S^-}^k$. For instance, in the discrete case this asks if it is possible to "flip" the classification of $\bar{x}$ by "flipping" at most $\ell$ of its components. Figure 2 illustrates how counterfactuals look in the continuous setting under the $\ell_2$-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 computation problems. In fact, as shown in the explainability literature, the hardness of a decision problem often implies hardness for its associated computation problem [Arenas et al.(2022)]. 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 $k \geqslant 1$. In the continuous case, hardness holds regardless of the norm being used.

**Theorem 1.** *The following statements hold:*

1. *$k$-MINIMUM SUFFICIENT REASON$(\{0, 1\}, D_H)$ is NP-hard under many-to-one reductions, for every fixed odd integer $k \geqslant 1$.*

2. *$k$-MINIMUM SUFFICIENT REASON$(\mathbb{R}, D_p)$ is NP-hard under many-to-one reductions for every fixed odd integer $k \geqslant 1$ and integer $p > 0$.*

*Proof.* We first consider the discrete case in item (1). We start by providing some intuition regarding our construction. Assume that $k = 1$. We reduce from the well-known NP-complete *Vertex Cover* problem: given an undirected graph $G = (V, E)$, and an integer $\ell \geqslant 0$, check whether there is a *vertex cover $C$* in $G$ of size $|C| \leqslant \ell$. Recall that a vertex cover is a subset of nodes $C \subseteq V$ such that every edge in $E$ has an endpoint in $C$. Let $G$ be a graph with vertex set $V = \{1, \ldots, n\}$ and edge set $E = \{e_1, \ldots, e_m\}$. Our encoding considers $n$-dimensional vectors (each component corresponds to a vertex). The vector $\bar{x}$ is the zero vector, which will be classified positively. The negative example set $S^-$ corresponds to the edges $e_1, \ldots, e_m$: each $e_j$ is encoded in the natural way, a component is 1 iff it is an endpoint of $e_j$. Hence, $S^-$ is simply an encoding of the incidence matrix of $G$. The positive example set $S^+$ contains all the vectors resulting from vectors in

$S^-$ by flipping a 1 to a 0. Then each vector in $S^-$ has two "guards" in $S^+$ that are closer to $\bar{x}$. Using these "guards", it follows that if $C$ is a vertex cover, then $C$ is a sufficient reason. On the other hand, if $C$ is not a vertex cover, an edge/vector $e_j$ not being covered directly provides a counterexample to $C$ being a sufficient reason. We formalize this reduction below.

Given an instance of Vertex Cover, we construct an instance of 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 $f_{S^+,S^-}^1(\bar{x}) = 1$, 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 $f_{S^+,S^-}^1$ 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 $f_{S^+,S^-}^1(\bar{z}) = 1$ 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

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

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 $f_{S^+,S^-}^1(\bar{y}_j) = 0$, which is a contradiction.

The hardness of $k$-Minimum Sufficient Reason($\{0, 1\}, D_H$), when $k \geq 3$, is deferred to Section 8, where an even stronger result is proved (Theorem 8).

Now we turn to item (2). The construction for the continuous case with $k \geq 1$ and $\ell_p$-norms is a direct adaptation of the previous idea. Now each of the vectors $e_j$ in $S^-$ must be represented by $(k+1)/2$ sufficiently close vectors: we increment the two components with value 1 by small quantities. In $S^+$ we add suitable "guards" as before.

Given an instance of Vertex Cover, we construct an instance of $k$-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 define $\bar{y}_{j,h}^1$ and $\bar{y}_{j,h}^2$ as the vectors obtained from $\bar{y}_{j,h}$ by changing the first and second components equal to $1 + \varepsilon_h$ to $\varepsilon_h$, respectively, and keeping all other components unchanged. We finally define:

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

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

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

Suppose first $C \subseteq \{1, \ldots, n\}$ is a vertex cover. We claim that $C$ is a sufficient reason for $\bar{x}$ with respect to $f_{S^+,S^-}^k$. 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

8

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 $f_{S^+,S^-}^k(\bar{z}) = 1$ 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 $\varepsilon_h$. Indeed, 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

$$\|\bar{z} - \bar{y}_{j,h}\|_p^p \;>\; \|\bar{z} - g(\bar{y}_{j,h})\|_p^p \quad\Longleftrightarrow\quad |\bar{z}[i] - (1 + \varepsilon_h)|^p \;>\; |\bar{z}[i] - \varepsilon_h|^p \quad\Longleftrightarrow\quad (1 + \varepsilon_h)^p \;>\; \varepsilon_h^p,$$

and then $g$ satisfies the required conditions.

Assume now that $X \subseteq \{1, \ldots, n\}$ is a sufficient reason. 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 $f_{S^+,S^-}^k(\bar{y}_{j,1}) = 0$, which is a contradiction. This follows from the fact that $\|\bar{y}_{j,1} - \bar{y}_{j,h}\|_p^p < \|\bar{y}_{j,1} - \bar{y}_{j',h'}^a\|_p^p$, for every $h \in \{1, \ldots, (k+1)/2\}$ and $\bar{y}_{j',h'}^a \in S^+$. Indeed, since $\bar{y}_{j',h'}^a$ has exactly one component with value $1 + \varepsilon_{h'}$, and $\bar{y}_{j,1}$ has two components with value $1 + \varepsilon_1$, we have:

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

This finishes the proof of the theorem. $\qquad\square$

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 the metric affects their tractability. To this end, we focus on two of the most commonly studied metrics: $\ell_2$ and $\ell_1$.

# 5   The Continuous Setting Based on the $\ell_2$-distance

It turns out that in case of the $\ell_2$-norm, all mentioned problems, apart from $k$-MINIMUM SUFFICIENT REASON, are tractable. The main reason is that in the case of the $\ell_2$-norm, an inequality $d_n(\bar{x}, \bar{a}) \leqslant d_n(\bar{x}, \bar{c})$ is the *linear* inequality in $\bar{x}$ given by $(\bar{a} - \bar{c})^\top \bar{x} \geqslant \frac{1}{2}(\bar{a} - \bar{c})^\top(\bar{a} + \bar{c})$. To see the correctness of this inequality, observe that with the $\ell_2$-norm, the set of equidistant points to $\bar{a}$ and $\bar{c}$ is a hyperplane $H$; see Figure 3. The vector $\bar{a} - \bar{c}$ is orthogonal to $H$, and thus $H$ is defined by the equality $(\bar{a} - \bar{c})^\top x = b$ for some $b$. To find $b$, we simply observe that the average point $\frac{1}{2}(\bar{a} + \bar{c})$ belongs to $H$, and hence $b = \frac{1}{2}(\bar{a} - \bar{c})^\top(\bar{a} + \bar{c})$. It is worth noticing that other norms behave differently, with the set of equidistant points not in a hyperplane (see Figure 4).

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, one for each pair of sets $A$ and $B$ in the statement of the proposition. This is a polynomial number of polyhedra 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 $\ell_2$-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 $k \geqslant 1$.*
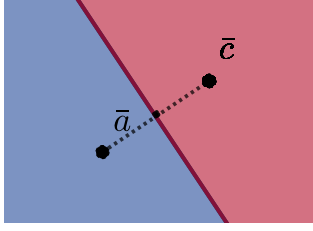
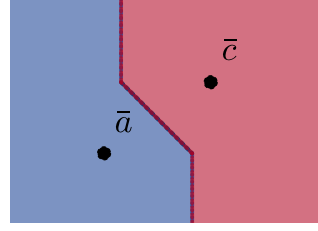Figure 3: Set of equidistant points between $\bar{a}$ and $\bar{c}$ for the $\ell_2$-norm.



Figure 4: Set of equidistant points between $\bar{a}$ and $\bar{c}$ for the $\ell_1$-norm.

*Proof.* Assume first that $f^k_{S^+,S^-}(\bar{x}) = 0$. Let $X \subseteq \{1, \ldots, n\}$ and consider the affine subspace

$$U(X, \bar{x}) := \{\bar{y} \in \mathbb{R}^n \mid \bar{x}[i] = \bar{y}[i], \text{ for every } i \in X\}.$$

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 intersection 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 [Schrijver(1998)].

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 $\varepsilon$, obtained by turning every strict inequality $l > 0$ of $S$ into a non-strict inequality $l \geqslant \varepsilon$. Feasibility of $S$ is equivalent to existence of a feasible solution to $\widehat{S}$ with positive $\varepsilon$. To find out if the latter is true, it is enough to find the optimal solution to the problem of maximizing $\varepsilon$ subject to $\widehat{S}$. $\square$

As a corollary, we obtain the following:

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

**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 [Kozlov et al.(1980)]. A very similar algorithm is described (although without a theoretical analysis) by Magagnini et al. [Magagnini et al.(2025)].

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

*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 $B_\ell(\bar{x}) = \{\bar{y} \in \mathbb{R}^n \mid \|\bar{y} - \bar{x}\|_2 \leqslant \ell\}$ 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 $\{\bar{y} \in \mathbb{R}^n \mid f^k_{S^+,S^-}(\bar{y}) = 1\}$, 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 [Kozlov et al.(1980)]. Specifically, we minimize $q(\bar{y}) = \|\bar{x} - \bar{y}\|^2_2$

10

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 $\ell^2$, with the optimal solution providing the counterfactual explanation.

Similarly, when $f_{S^+,S^-}^k(\bar{x}) = 1$, 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 [Kozlov et al.(1980)] 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 [Kozlov et al.(1980)]. 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 $\ell^2$.

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 [Hiriart-Urruty and Lemaréchal(2001)], 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 $\ell_2$-distances are used.

**Corollary 2.** *Consider the setting $(\mathbb{R}, D_2)$ and let $k \geqslant 1$ 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 $f_{S^+,S^-}^k$ at distance at most $\ell$ in case there exists one.*

*Proof.* This reduces to finding a point in the intersection of the ball $B_\ell(\bar{x})$ and an open polyhedron $P$, if this intersection is non-empty. If it is non-empty, then $\widehat{P}$ intersects the interior of $B_\ell(\bar{x})$. We can find a point $\bar{y}$ in this intersection by minimizing the quadratic form $q(\bar{y}) = \|\bar{x} - \bar{y}\|_2^2$ subject to $\widehat{P}$. Suppose $\bar{y}$ lies on the boundary of $\widehat{P}$ (otherwise, we are done). 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 $\beta$ such that $\langle \alpha, \beta \rangle > 0$ for every inequality $\langle \alpha, \bar{y} \rangle \geqslant c$ which turns into equality on $\bar{y}$. After such $\beta$ is found, we just need to move from $\bar{y}$ along $\beta$ a little bit so that we still inside $B_\ell(\bar{x})$. Finding such $\beta$ 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. □

**The parameterized complexity of Counterfactual Explanation for the $\ell_2$-distance.** Running time of our algorithms for problems $k$-Minimal Sufficient Reason($\mathbb{R}, D_2$) and $k$-Counterfactual Explanation($\mathbb{R}, D_2$) depends exponentially on $k$. More specifically, it is $n^{O(k)}$, where $n$ is the size of the dataset. This comes from the fact that in our algorithms we essentially go through all possible $k$-sized sets of points of the dataset that could be $k$ closest neighbors.

This raises a question – is there a fixed-parameter tractable algorithm for our problems in the case of $\ell_2$-distance, where $k$ is considered as a parameter? That is, are there algorithms running in time $f(k) \cdot n^c$, for some computable function $f$ and constant $c$? We show that this is unlikely for the counterfactual explanations.

**Theorem 3.** *The problem $k$-Counterfactual Explanation($\mathbb{R}, D_2$) is W[1]-hard [2] when $k$ is a parameter, and NP-complete when $k$ is part of the input.*

*Proof.* Both results follow from the following lemma.

**Lemma 1.** *There exists a polynomial-time many-to-one reduction from the $k$-RegClique problem (given a regular graph $G$, decide, whether it has a $k$-clique) to the $(2k-1)$-Counterfactual Explanation($\mathbb{R}, D_2$).*

Indeed, $k$-RegClique is known to be both W[1]-complete [Cygan et al.(2015), Theorem 13.4] when $k$ is a parameter and NP-complete when $k$ is a part of the input [Garey and Johnson(1979), Problem GT.20, page 194].

In the rest of the proof, we establish Lemma 1 (the fact that $k$-Counterfactual Explanation($\mathbb{R}, D_2$) belongs to NP is immediate). For readability, throughout the proof we omit the bar notation and write vectors as $x, x_i, y, 0, \ldots$. For our reduction, we first require the following lemma that provides an embedding of the nodes of our graph into an Euclidean space that respects the connections between the nodes. The resulting vectors are actually $0, 1$-vectors so we formulate our lemma for the Hamming distance.

**Lemma 2.** *Let $G$ be a regular graph with $n$ nodes and degree $d$. Let $m = n^2 + n + d - 5$. In $O(n^3)$-time, we can assign to each node of $G$ a vector in $\{0,1\}^m$ such that:*

   a) *all $n$ vectors have Hamming weight $2(n + d - 3)$;*

   b) *vectors that correspond to any two distinct nodes of $G$ that are connected by an edge have Hamming distance $2(n + d - 3)$;*

   c) *vectors that correspond to any two distinct nodes of $G$ that are not connected by an edge have Hamming distance $2(n + d - 1)$.*

*Proof.* For simplicity, let us first explain how to satisfy just conditions b) and c). We can do this using $n^2$ coordinates that will be split into $n$ blocks of size $n$. Both the blocks and coordinates in each block are indexed by the nodes of $G$.

We define the vector of a node $u$ as follows. In each of its blocks that corresponds to a vertex, different from $u$, we put the "one-hot encoding" of $u$ (the vector that has 1 in the $u$-coordinate and 0 in the rest of the coordinates). Now, in the $u$-block we put the vector that has 1s exactly in coordinates that correspond to neighbors of $u$ (thus, it will have $d$ ones).

Take any two distinct nodes $u, v$ and their corresponding vectors $f_u, f_v$. In the blocks that correspond to vertices, distinct from both $u$ and $v$, both $f_u$ and $f_v$ have a single 1 in distinct position. These blocks thus contribute $2(n - 2)$ to the Hamming distance. Now, consider a $u$-block or a $v$-block. In these blocks, one of the vectors $f_u, f_v$ have $d$ ones, and the other just a single one. If $u, v$ are connected in $G$, then both blocks have a common one so that they contribute $2(d - 1)$ to the Hamming distance. Now, if $u, v$ are not connected, in the $u$-block and $v$-block with have $d$ ones and 1 separate one, contributing $2(d + 1)$ to the Hamming distance. This establishes conditions b) and c).

To obtain condition a), observe that, in the current construction, all vectors have the same number of 1s, namely $n - 1 + d$. It thus suffices to add $2(n + d - 3) - (n - 1 + d) = n + d - 5$ coordinates where all vectors have 1s. $\square$

We now explain how we get from a $d$-regular graph $G$ to an instance of the $(2k-1)$-Counterfactual Explanation($\mathbb{R}, D_2$) problem. We first give an argument for the case when points in a dataset can have multiplicities. In the end we will explain how to get rid of this assumption.

Namely, we take $0, 1$-vectors from Lemma 2, assigned to nodes of $G$, all classified positively. We also add the all-0 vector to the dataset, classified negatively and with multiplicity $k$. Due to conditions a), b), c), the all-0 point has the $\ell_2$-distance $\alpha = \sqrt{2(n + d - 3)}$ from all the other points of the dataset. Any two positive points of the data set that correspond to two connected nodes of $G$ also have the $\ell_2$-distance equal

---

[2]The class W[1] is the analog of NP in parameterized complexity. For further details see [Cygan et al.(2015)].

to $\alpha$ between them. At the same time, points, corresponding to nodes, not connected by an edge, are all at distance $\beta = \sqrt{2(n + d - 1)}$ from each other.

In our instance of the $(2k - 1)$-Counterfactual Explanation$(\mathbb{R}, D_2)$ problem, the question will be: is there a point at distance at most $R$ (a number that will be defined later) from the all-0 vector $x = (0, \ldots, 0)$ where $f^{2k-1}_{S^+, S^-}$ takes the positive value? Observe that at $x$, the $(2k - 1)$-NN takes the negative value – precisely at 0, there is a negative point with multiplicity $k$, and positive points are at some positive distance from 0.

As the point 0 has multiplicity $k$, in order for $f^{2k-1}_{S^+, S^-}(y)$ to be positive for a point $y$, there have to exist $k$ positive points $x_1, \ldots, x_k$ in the dataset such that $y$ is closer or at the same distance to all these points as to 0. The minimal distance from 0 at which such $y$ exists is formally defined in the following definition.

**Definition 1.** *Let $x_1, \ldots, x_k$ be $k$ points in an Euclidean space $\mathbb{R}^m$ for some $m$. Define $r(x_1, \ldots, x_k)$ to be the minimum of $\|y\|_2$ over all $y \in \mathbb{R}^m$ satisfying:*

$$\|y - x_1\|_2 \leqslant \|y\|_2, \ldots, \|y - x_k\|_2 \leqslant \|y\|_2. \tag{1}$$

*(if no such $y$ exists, we set $r(x_1, \ldots, x_k) = +\infty$).*

The answer to the counterfactual explanation problem is thus determined by whether there exist $k$ positive points $x_1, \ldots, x_k$ in the dataset with $r(x_1, \ldots, x_k) \leqslant R$. We now relate this with the existence of the $k$-clique in the initial graph $G$.

If a $k$-clique exists, the corresponding $k$ points $x_1, \ldots, x_k$ of the dataset will all be at distance $\alpha$ from each other and from 0. As we will see later, $r(x_1, \ldots, x_k) = \alpha \sqrt{\frac{k}{2(k+1)}}$ in this case. Now, when no $k$-clique exists, we will argue that $r(x_1, \ldots, x_k)$ is slightly bigger than $\alpha \sqrt{\frac{k}{2(k+1)}}$ for any $k$ positive points $x_1, \ldots, x_k$. This will happen because 0 is still at distance $\alpha$ from $x_1, \ldots, x_k$, and distances between $x_1, \ldots, x_k$ are all at least $\alpha$ with at least one distance being at least $\beta > \alpha$ (for any $k$ nodes there will be at least two nodes not connected by an edge, so their points will be at a larger distance). The following lemma justifies this reasoning.

**Lemma 3.** *Let $x_1, \ldots, x_k$ be $k$ points in an Euclidean space $\mathbb{R}^m$ for some $m$. Further, let $0 < \alpha \leqslant \beta$. Then the following hold:*

a) *if the $\ell_2$-distance between any two points in $\{0, x_1, \ldots, x_k\}$ is precisely $\alpha$, then*

$$r(x_1, \ldots, x_k) \leqslant \alpha \sqrt{\frac{k}{2(k + 1)}}$$

b) *if the $\ell_2$-distance from 0 to any point in $\{x_1, \ldots, x_k\}$ is precisely $\alpha$, the distance between any two points in $\{x_1, \ldots, x_k\}$ is at least $\alpha$, and for at least one pair of points in $\{x_1, \ldots, x_k\}$ the distance is at least $\beta$, then*

$$r(x_1, \ldots, x_k) \geqslant \alpha \sqrt{\frac{k}{2(k + 1 - \delta)}}, \qquad where \ \delta = \frac{\beta^2 - \alpha^2}{k\alpha^2}.$$

*Proof.* Within the proof, we omit the subscript $\|\|_2$ in the notation for the $\ell_2$-norm.

Let us start with a). For any $i, j \in \{1, \ldots, k\}$, we have:

$$\langle x_i, x_j \rangle = \frac{\langle x_i, x_i \rangle + \langle x_j, x_j \rangle - \langle x_i - x_j, x_i - x_j \rangle}{2} = \begin{cases} \alpha^2 & i = j, \\ \alpha^2/2 & i \neq j. \end{cases}$$

We take $y = (x_1 + \ldots + x_k)/(k + 1)$, the center of mass of the regular simplex, formed by $0, x_1, \ldots, x_k$. It is of equal distance to all the points $0, x_1, \ldots, x_k$, thus satisfying (1). The square of this distance is computed as:

$$\langle \frac{x_1 + \ldots + x_k}{k + 1}, \frac{x_1 + \ldots + x_k}{k + 1} \rangle = \frac{k\alpha^2 + (k^2 - k)\alpha^2/2}{(k + 1)^2} = \frac{k\alpha^2}{2(k + 1)},$$

giving us the required upper bound.

We now establish b). Let $y$ be any point, satisfying (1). Rewriting these conditions in terms of dot-products, we obtain

$$\langle y - x_i, y - x_i \rangle \leqslant \langle y, y \rangle \iff \langle y, x_i \rangle \geqslant \|x_i\|^2/2 = \alpha^2/2 \qquad i = 1, \ldots, k. \tag{2}$$

Define $z = (x_1 + \ldots + x_k)$. We lower bound $\|y\|$ by the length of the projection of the vector $y$ to the direction of the vector $z$, that is,

$$\|y\| \geqslant \frac{|\langle y, z \rangle|}{\|z\|}.$$

Let us first explain why $z \neq 0$. Indeed, by (2), the dot-product $\langle y, x_i \rangle$ is strictly positive for every $i = 1, \ldots, k$ as $\|x_i\| \geqslant \alpha > 0$. Hence, $\langle y, z \rangle$ is also strictly positive as a sum of positive values. As a result, we obtain:

$$\|y\| \geqslant \frac{|\langle y, z \rangle|}{\|z\|} = \frac{\langle y, x_1 \rangle + \ldots + \langle y, x_k \rangle}{\sqrt{\langle x_1 + \ldots + x_k, x_1 + \ldots + x_k \rangle}}$$

The numerator in the last expression is bounded from below by $k\alpha^2/2$ due to (2). We now upper bound the denominator. The expression $\langle x_1 + \ldots + x_k, x_1 + \ldots + x_k \rangle$ has $k$ terms of the form $\|x_i\|^2$ for $i = 1, \ldots, k$ that contribute precisely $k\alpha^2$ to the sum. Additionally, we have $k^2 - k$ dot-products of the form $\langle x_i, x_j \rangle$ for $i \neq j$. Each of them can be written as

$$\langle x_i, x_j \rangle = \frac{\langle x_i, x_i \rangle + \langle x_j, x_j \rangle - \langle x_i - x_j, x_i - x_j \rangle}{2} = \alpha^2 - \|x_i - x_j\|^2/2.$$

By the conditions of the lemma, each of these products is bounded from above by $\alpha^2/2$, but for at least one of them, the upper bound is even better, namely $\alpha^2 - \beta^2/2$, that is, smaller by $(\beta^2 - \alpha^2)/2$. We thus can upper bound the product $\langle x_1 + \ldots + x_k, x_1 + \ldots + x_k \rangle$ as follows:

$$\langle x_1 + \ldots + x_k, x_1 + \ldots + x_k \rangle \leqslant k\alpha^2 + (k^2 - k)\alpha^2/2 - (\beta^2 - \alpha^2)/2.$$

Overall, we obtain the following lower bound on $\|y\|$:

$$\|y\| \geqslant \frac{k\alpha^2/2}{\sqrt{k\alpha^2 + (k^2 - k)\alpha^2/2 - (\beta^2 - \alpha^2)/2}} = \alpha \cdot \sqrt{\frac{k \cdot k\alpha^2}{2 \cdot 2 \cdot (k\alpha^2 + (k^2 - k)\alpha^2/2 - (\beta^2 - \alpha^2)/2)}}$$

$$= \alpha \sqrt{\frac{k \cdot k\alpha^2}{2 \cdot ((k+1)k\alpha^2 - (\beta^2 - \alpha^2))}} = \alpha \sqrt{\frac{k}{2(k+1) - \frac{2(\beta^2 - \alpha^2)}{k\alpha^2}}},$$

as required. □

The above reduction clearly takes polynomial time, but it remains to pick a rational number $R$, satisfying:

$$\lambda_1 = \alpha \sqrt{\frac{k}{2(k+1)}} \leqslant R < \alpha \sqrt{\frac{k}{2(k+1-\delta)}} = \lambda_2, \qquad \delta = \frac{\beta^2 - \alpha^2}{k\alpha^2}. \tag{3}$$

To avoid technicalities, we slightly modify the construction so that $\lambda_1$ becomes a rational number and we can set $R = \lambda_1$. Namely, just repeat every coordinate $T$ times in all vectors of the dataset, for $T$ to be defined later. All relative distances will be multiplied by $\sqrt{T}$. After that, the value of $\lambda_1$ will become:

$$\lambda_1 = \sqrt{T \cdot 2(n + d - 3)} \sqrt{\frac{k}{2(k+1)}} = \sqrt{(n + d - 3)T \frac{k}{k+1}}.$$

Choosing $T = (n + d - 3)k(k + 1)$, we get

$$\lambda_1 = (n + d - 3)k \tag{4}$$

We now explain how to get rid of the multiplicities in our dataset. Instead of just point $0$ with multiplicity $k$, we will have $0$ and $k - 1$ other negative points $y_1, \ldots, y_{k-1}$ with the following properties:

14

- every $y_i$ is at the same distance $\alpha$ from all positive points of the dataset.

- every $y_i$ is $\varepsilon$-close to 0, for some $\varepsilon > 0$ small enough that $R + \varepsilon < \lambda_2$.

Assuming we have constructed such $y_1, \ldots, y_{k-1}$, we explain why the reduction still works after this modification. Indeed, in order to change the $(2k-1)$-NN classification, we have to become closer (or at the same distance) to some $k$ positive points as to some negative points. If a $k$-clique exists, we can take $k$ corresponding positive points $x_1, \ldots, x_k$, we can become as close to them as to 0 by moving at distance at most $R = \alpha \sqrt{\frac{k}{2(k+1)}}$ (as before, the points $0, x_1, \ldots, x_k$ form a configuration as in the item a) of Lemma 3).

Now, assume that no $k$-clique exists. Pick arbitrary $k$ positive points $x_1, \ldots, x_k$. By the same argument as before, we require distance at least $\lambda_2 > R$ to become at least as close to $x_1, \ldots, x_k$ as to 0. However, we should also argue that we cannot become as close to $x_1, \ldots, x_k$ as to $y_i$ for some $i$, by moving at distance at most $R$. Indeed, $x_1, \ldots, x_k$ correspond to a non-clique, and $y_i$ has distance $\alpha$ to all of them, so $y_i, x_1, \ldots, x_k$ form the same configuration as described in item b) of Lemma 3 (after translation by $-y_i$). Hence, we have to move at least $\lambda_2$ from $y_i$, and hence, at least $\lambda_2 - \varepsilon > R$ from 0 (because 0 and $y_i$ are $\varepsilon$-close).

It remains to construct $y_1, \ldots, y_{k-1}$ with the required properties. Let us first calculate the value of $\varepsilon$ that would work. Recall that $\alpha = \sqrt{T \cdot 2(n+d-3)}, \beta = \sqrt{T \cdot 2(n+d-1)}$, so we get the following expression for $\delta$ in (3):

$$\delta = \frac{2}{k \cdot (n+d-3)}.$$

Define

$$\lambda_2' = \alpha \sqrt{\frac{k}{2(k+1-\delta')}}, \qquad \text{where } \delta' = \left(1 - \left(\frac{n^4-1}{n^4}\right)^2\right)(k+1)$$

Observe that

$$0 < \delta' = \left(1 - \frac{n^4-1}{n^4}\right)\left(1 + \frac{n^4-1}{n^4}\right)(k+1) \leqslant 2(k+1)/n^4 < \frac{2}{k \cdot (n+d-3)} = \delta,$$

where the last inequality is due to the fact that $k, d \leqslant n$, meaning that $\lambda_1 < \lambda_2' < \lambda_2$. The value of $\delta'$ is also chosen so that $\lambda_2'$ is rational. Indeed, $\lambda_1$ is rational, and the fraction $\lambda_1/\lambda_2'$ is equal to:

$$\sqrt{\frac{k+1-\delta'}{k+1}} = \frac{n^4-1}{n^4}.$$

Recalling (4), we get that it suffices to set $\varepsilon = \frac{1}{n^4} < \lambda_2' - \lambda_1 = \frac{k(n+d-3)}{n^4-1}$.

We now construct $y_1, \ldots, y_{k-1}$. Recall also that all positive points were 0,1-vectors with the same number of 1s $W$. Let $m$ be the dimensionality of vectors in our dataset. It is equal to the dimensionality of vectors from Lemma 2 times $T$. In the end of the argument, we will require a simple bound $m \geqslant n$. Define

$$\tau = \frac{1}{m^{100}}, \qquad S = 2Wm^{100} - m.$$

We add $S(k-1)$ new coordinates where all previous points of the dataset have 0s. Each point $y_1, \ldots, y_{k-1}$ will have value $\tau$ in all $m$ old coordinates. Now, we split new coordinates into $k-1$ blocks of size $S$. Each $y_i$ for $i = 1, \ldots, k-1$ will additionally have $S$ coordinates of the $i$th block equal to $\tau$, with the rest of the new coordinates equal to 0.

We need to show two things: (a) all the points $y_1, \ldots, y_{k-1}$ are $\varepsilon$-close to 0; (b) all of them have the same distance to all positive points of the dataset as 0. We start with (a). Each $y_1, \ldots, y_{k-1}$ has $m + S$ coordinates, equal to $\tau$, so the square of its distance to 0 is:

$$(m+S)\tau^2 = \frac{2Wm^{100}}{m^{200}} \leqslant \frac{2}{m^{99}} < \frac{1}{n^8} = \varepsilon^2,$$

15

because $m \geqslant n$.

We now show (b). As we have mentioned, every positive point $x_j$ of the dataset has $W$ 1s in the old $m$ coordinates, and just 0s in the new coordinates. Hence, square of the distance from 0 to it is $W$. The following calculation shows that the square of the distance from each $y_i$ to $x_j$ is also $W$, thus finishing the proof of the theorem:

$$\|y_i - x_j\|^2 = (m - W)\tau^2 + W(1 - \tau)^2 + S\tau^2 = m\tau^2 + W - 2W\tau + S\tau^2 = W + (m + S)\tau^2 - 2W\tau$$

$$= W + \frac{2Wm^{100}}{m^{200}} - \frac{2W}{m^{100}} = W. \; \square$$

# 6 The Continuous Setting Based on the $\ell_1$-distance

We first show that the positive results for counterfactual explanations under the $\ell_2$-norm do not extend to the $\ell_1$-norm. For abductive explanations, we demonstrate that some favorable properties of the $\ell_2$-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 $\ell_1$-distance is NP-hard even when $S^+$ and $S^-$ are of minimal non-degenerate size.

**Theorem 4.** *For every odd integer $k \geqslant 1$, 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 linear program, constructable in polynomial time from the input. We now show the NP lower bound. Let us start with the case $k = 1$ (we reduce from $k = 1$ to the general case in the end of the proof). 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 $\sum_{i \in T} w_i \leqslant W$ and $\sum_{i \in T} v_i \geqslant (v_1 + \ldots + v_n)/2$. The hardness of this problem arises from a reduction from the partition problem [Lewis and Papadimitriou(1997)].

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 $\gamma \cdot v_i \leqslant 1/2 < 1$ 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 $\|\bar{h} - \bar{0}\|_1 < \|\bar{g} - \bar{0}\|_1$, resulting in $f^1_{S^+, S^-}(\bar{0}) = 0$. We are asked if there exists $\bar{y}$ with $\|y\|_1 \leqslant \ell = W$ such that $f^1_{S^+, S^-}(\bar{y}) = 1$, or, equivalently, $\|\bar{h} - \bar{y}\|_1 \geqslant \|\bar{g} - \bar{y}\|_1$. 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 $\|\bar{h} - \bar{y}\|_1 \geqslant \|\bar{g} - \bar{y}\|_1$, or equivalently:

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

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

$$\sum_{i=1}^{n} (|\bar{h}_i - \bar{y}_i| - |\bar{g}_i - \bar{y}_i| + \gamma v_i)/2 \; \geqslant \; \gamma(v_1 + \ldots + v_n)/2. \tag{5}$$

16

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 $\gamma v_i$. Thus, if $\bar{y}_i = 0$, the left-hand side of (5) contributes 0; if $\bar{y}_i = \bar{g}_i$ (the right endpoint), it contributes $\gamma v_i$. Therefore, the left-hand side of (5) is the sum of the values of the items in the knapsack, scaled by $\gamma$. This proves (5), 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 $\|\bar{y}\|_1 \leqslant \ell = W$ and $\|\bar{h} - \bar{y}\|_1 \geqslant \|\bar{g} - \bar{y}\|_1$, with the latter being equivalent to (5). 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, $\bar{y}_i \leqslant \bar{g}_i$ for every $i = 1, \dots, n$, as otherwise we can decrease $\|\bar{y}\|_1$ without changing the left-hand side of (5). 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 $\bar{y}_i$ lies to the left of $[\bar{h}_i, \bar{g}_i]$, then $\bar{y}_i = 0$ for every $i = 1, \dots, n$. Again, if not, $\|\bar{y}\|_1$ can be decreased without changing the left-hand side of (5).

We now establish that, without loss of generality, we may assume $\bar{y}_i = 0$ or $\bar{y}_i = \bar{g}_i$ for every $i = 1, \dots, n$. 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 (5) will begin to decrease and increase, respectively, at twice the same rate, leaving the left-hand side of (5) 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 (5). 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 (5), so the corresponding inequality is still satisfied. We now explain why $\|\bar{y}\|_1 \leqslant W = \ell$ 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, $\|\bar{y}\|_1 = |\bar{y}_1| + \dots + |\bar{y}_n|$ 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 $i = 1, \dots, n$. We place objects with $\bar{y}_i = w_i$ into the knapsack. The condition $\|\bar{y}\|_1 \leqslant \ell = W$ 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 (5) becomes the sum of $\gamma v_i$ 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.

We only have to show that it is enough to establish the theorem for $k = 1$. This is because 1-Counterfactual Explanation$(\mathbb{R}, D_1)$ for $|S^+| = |S^-| = 1$ reduces to $k$-Counterfactual Explanation$(\mathbb{R}, D_1)$ for $|S^+| = |S^-| = (k+1)/2$ for every odd integer $k \geqslant 1$. Indeed, consider any input to 1-Counterfactual Explanation$(\mathbb{R}, D_1)$ with $|S^+| = |S^-| = 1$. It consists of two singleton sets $S^+ = \{\bar{g}\}$ and $S^- = \{\bar{h}\}$ for $\bar{g}, \bar{h} \in \mathbb{R}^n$, a vector $\bar{x} \in \mathbb{R}^n$ and an integer $\ell > 0$. The question is whether there exists $\bar{y} \in \mathbb{R}^n$ such that $\|\bar{x} - \bar{y}\|_1 \leqslant \ell$ and $f^1_{S^+, S^-}(\bar{x}) \neq f^1_{S^+, S^-}(\bar{y})$. In this special case, we have $f^1_{S^+, S^-}(\bar{z}) = 1$ if $\|\bar{z} - \bar{g}\|_1 \leqslant \|\bar{z} - \bar{h}\|_1$ and $f^1_{S^+, S^-}(\bar{z}) = 0$ otherwise. In other words, $f^1_{S^+, S^-}(\bar{z})$ can be defined as the sign of the difference $\|\bar{z} - \bar{h}\|_1 - \|\bar{z} - \bar{g}\|_1$ (where the sign of a non-negative number is 1 and the sign of a negative number is 0).

Without loss of generality, we may assume that $\bar{x} = \bar{0}$. We now transform our input into an input for $k$-Counterfactual Explanation$(\mathbb{R}, D_1)$ with $|S^+| = |S^-| = (k+1)/2$. We first add arbitrary $(k-1)/2$ points to $S^+$ and $(k-1)/2$ points to $S^-$, for example:

$$\bar{p}_1 = (1, 0, \dots, 0), \ \dots, \ \bar{p}_{\frac{k-1}{2}} = \left(\frac{k-1}{2}, 0, \dots, 0\right) \text{ to } S^+,$$

$$\bar{p}_{\frac{k+1}{2}} = \left(\frac{k+1}{2}, 0, \dots, 0\right), \ \dots, \ \bar{p}_{k-1} = (k-1, 0, \dots, 0) \text{ to } S^-.$$

We then add one more coordinate to all vectors, where we put 0 for $\bar{x}, \bar{p}_1, \ldots, \bar{p}_{k-1}$ and $M = 10(\ell + k)$ for $\bar{g}, \bar{h}$. Let $B_\ell$ be the ball of radius $\ell$ around $\bar{x} = \bar{0}$ (in the space with the new coordinate). Observe that any point among $\bar{p}_1, \ldots, \bar{p}_{k-1}$ is much closer to any point in $B_\ell$ than $\bar{g}$ and $\bar{h}$. Hence, for any potential counterfactual explanation $\bar{y} \in B_\ell$, the closest $k-1$ points in $S^+ \cup S^-$ will be $\bar{p}_1, \ldots, \bar{p}_{k-1}$, among which half is from $S^+$ and half is from $S^-$. Hence, the value $f^k_{S^+,S^-}(\bar{y})$ for any $\bar{y} \in B_\ell$ is determined by whether $\bar{g}$ or $\bar{h}$ (with an additional coordinate) is closer to $\bar{y}$, with value 1 in case of the equality. This means that the answer to the problem is determined by whether the sign of $\|\bar{y} - \bar{h}\|_1 - \|\bar{y} - \bar{g}\|_1$ is different for some $\bar{y} \in B_\ell$ from the sign of $\|\bar{x} - \bar{h}\|_1 - \|\bar{x} - \bar{g}\|_1$. Despite we have an additional coordinate, the answer to this is the same as in the initial 1-Counterfactual Explanation problem. This is because $\bar{g}$ and $\bar{h}$ have the same value in this additional coordinate so that it cannot contribute to the difference, meaning that we can restrict ourselves to $\bar{y} \in B_\ell$ that have 0 in this additional coordinate. □

**Abductive explanations**  Next, we observe that the 1-Check Sufficient Reason problem is tractable for the $\ell_1$-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:

$$\|(\bar{x}_1, \bar{y}_2) - (\bar{a}_1, \bar{a}_2)\|_1 \leqslant \|(\bar{x}_1, \bar{y}_2) - (\bar{c}_1, \bar{c}_2)\|_1 \text{ for every } \bar{c} \in S^-. \tag{6}$$

Using linearity of the $\ell_1$-norm under concatenation of vectors, we rewrite (6) as follows:

$$\|\bar{x}_1 - \bar{a}_1\|_1 - \|\bar{x}_1 - \bar{c}_1\|_1 \leqslant \|\bar{y}_2 - \bar{c}_2\|_1 - \|\bar{y}_2 - \bar{a}_2\|_1 \text{ for every } \bar{c} \in S^-. \tag{7}$$

The left-hand side of (7) does not depend on $\bar{y}_2$. The right-hand side of (7), for every $\bar{c}_2$, attains its maximum at $\bar{y}_2 = \bar{a}_2$, by the triangle inequality. This gives the following algorithm for our problem: for every $\bar{a} \in S^+$, check if $\bar{y}_2 = \bar{a}_2$ satisfies all inequalities in (6). 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* $X$ *for* $\bar{x}$ *with respect to* $f^1_{S^+,S^-}$.

We finish this section by showing that, starting from $k = 3$, the problems $k$-Check Sufficient Reason$(\mathbb{R}, D_1)$ and $k$-Minimal Sufficient Reason$(\mathbb{R}, D_1)$ become intractable.

**Theorem 5.** *For every odd integer* $k \geqslant 3$, *the problem* $k$-Check Sufficient Reason$(\mathbb{R}, D_1)$ *is* coNP-*complete, and the problem* $k$-Minimal Sufficient Reason$(\mathbb{R}, D_1)$ *is* NP-*hard under Turing reductions.*

*Proof.* Clearly, the $k$-Check Sufficient Reason$(\mathbb{R}, D_1)$ is in coNP. Indeed, if a given set of coordinates is not a sufficient reason, then a counter-example can be found as a solution to a linear program, constructible from the input in polynomial time.

We now show the coNP-hardness of the $k$-Check Sufficient Reason$(\mathbb{R}, D_1)$ problem. We reduce from the complement to the partition problem, where we are given $n$ positive integers $v_1, \ldots, v_n$, and the question is whether there exists $T \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in T} v_i = \sum_{i \notin T} v_i$. NP-completeness of the partition problem (as well as coNP-completeness of its complement) is standard [Lewis and Papadimitriou(1997)].

Let us first assume for simplicity that the same point can appear in the data set multiple times (we will explain later how to get rid of this simplification). Given $n$ positive integers $v_1, \ldots, v_n$ (an input to the partition problem), consider an $n$-dimensional data set $S$, consisting of the points

$$\bar{\alpha} = (0, 0, \ldots, 0) \qquad \text{classified 1, multiplicity 1,}$$
$$\bar{\beta} = (2v_1, 2v_2, \ldots, 2v_n) \qquad \text{classified 1, multiplicity } (k-1)/2,$$
$$\bar{\gamma} = (v_1, v_2, \ldots, v_n) \qquad \text{classified 0, multiplicity } (k+1)/2.$$

The size of the dataset is $k + 1$, and we have a point $\bar{\gamma}$ with multiplicity $(k+1)/2$, classified as 0. Hence, we have $f^k_{S^+, S^-}(\bar{x}) = 0$ if and only if $\bar{\gamma}$ is not the farthest point to $\bar{x}$ in the dataset, i.e., if $\|\bar{\gamma} - \bar{x}\|_1 < \|\bar{\alpha} - \bar{x}\|_1$ or $\|\bar{\gamma} - \bar{x}\|_1 < \|\bar{\beta} - \bar{x}\|_1$ (here we require strict inequalities because of our optimistic view on $k$-NN classification). Observe that $f^k_{S^+, S^-}(\bar{\alpha}) = 0$ because $\|\bar{\gamma}\|_1 < \|\bar{\beta}\|_1$. We claim that there exists $\bar{y} \in \mathbb{R}^n$ such that $f^k_{S^+, S^-}(\bar{y}) = 1$ if and only if the initial partition problem has a solution. In other words, we claim that the empty set of coordinates is *not* a sufficient reason for $\bar{x} = \bar{0}$ if and only if the initial partition problem has a solution.

The condition $f^k_{S^+, S^-}(\bar{y}) = 1$ can be rewritten as

$$\begin{cases} \|\bar{\alpha} - \bar{y}\|_1 \leqslant \|\bar{\gamma} - \bar{y}\|_1 \\ \|\bar{\beta} - \bar{y}\|_1 \leqslant \|\bar{\gamma} - \bar{y}\|_1 \end{cases},$$

or, equivalently,

$$\begin{cases} \sum_{i=1}^{n} (|\bar{y}_i - 0| - |\bar{y}_i - v_i|) \leqslant 0 \\ \sum_{i=1}^{n} (|\bar{y}_i - 2v_i| - |\bar{y}_i - v_i|) \leqslant 0. \end{cases}$$

We claim that we can restrict our search to $\bar{y} \in \mathbb{R}^n$ where $\bar{y}_i \in \{0, 2v_i\}$ for every $i = 1, \ldots, n$. Indeed, consider the two expressions that appear in our conditions as functions of $\bar{y}_i$:

$$A_1(\bar{y}_i) = |\bar{y}_i - 0| - |\bar{y}_i - v_i|, \qquad A_2(\bar{y}_i) = |\bar{y}_i - 2v_i| - |\bar{y}_i - v_i|.$$

If $\bar{y}_i$ is to the left of $v_i$, then we can decrease $A_1$ without changing $A_2$ by moving $\bar{y}_i$ to 0. In turn, if $\bar{y}_i$ is to the right of $v_i$, we can decrease $A_2$ without changing $A_1$ by moving $\bar{y}_i$ to $2v_i$.

Now, take any "restricted" $\bar{y}$. Let $T$ denote the set of $i \in \{1, \ldots, n\}$ such that $\bar{y}_i = 2v_i$ (for $i \notin T$, we have $\bar{y}_i = 0$). Then our conditions can be rewritten as:

$$\begin{cases} \sum_{i \in T}^{n} v_i + \sum_{i \notin T}^{n} (-v_i) \leqslant 0 \\ \sum_{i \in T}^{n} (-v_i) + \sum_{i \notin T}^{n} v_i \leqslant 0. \end{cases}$$

These conditions are equivalent to $\sum_{i \in T}^{n} v_i = \sum_{i \notin T}^{n} v_i$, as required in the partition problem.

We now explain how to get rid of points with multiplicities. There are $k + 1$ points in our dataset (taking into account multiplicities). We add $k + 1$ auxiliary coordinates. For $i = 1, \ldots, k + 1$, for the $i$th point of the dataset, we put 1 to the $i$-th auxiliary coordinate and 0 to all the other auxiliary coordinates. For example, for $k = 3$, we obtain the following dataset:

$$\begin{array}{ll} (1, 0, 0, 0, \bar{\alpha}) & \text{classified 1,} \\ (0, 1, 0, 0, \bar{\beta}) & \text{classified 1,} \\ (0, 0, 1, 0, \bar{\gamma}) & \text{classified 0,} \\ (0, 0, 0, 1, \bar{\gamma}) & \text{classified 0.} \end{array}$$

We consider $\bar{x} = (0, \ldots, 0)$ as the input and ask, whether the set of auxiliary coordinates is a sufficient reason for $\bar{x}$. The same argument as before shows that this is the case if and only if the initial instance of the partition problem has a solution. Indeed, we are not allowed to change 0s in the $k + 1$ auxiliary coordinates of $\bar{x}$. Hence, these coordinates contribute 1 to distances from $\bar{y}$ to all the points. This does not change the differences of distances on which our argument is based on.

Finally, we explain why this reduction implies that $k$-Minimal Sufficient Reason($\mathbb{R}, D_1$) is NP-hard (under Turing reductions) as well. The crucial thing is that the set of $X$ for which we ask, whether $X$ is a sufficient reason, has constant-size, namely, $k + 1$. It remains to note that $X$ is a sufficient reason if and only if it has a subset which is a minimal sufficient reason. Hence, if we could solve $k$-Minimal Sufficient Reason($\mathbb{R}, D_1$) in polynomial-time, we could solve the partition problem in polynomial-time by running the algorithm for the minimal sufficient reason problem on all subsets of $X$. □

# 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 $\ell_1$-distance; however, our proof techniques differ.

**Counterfactual explanations**  We prove that $k$-Counterfactual Explanation is intractable for every odd integer $k \geqslant 1$, employing different techniques than those used for the NP-hardness of the problem in the continuous setting under the $\ell_1$-distance. Unlike the continuous case, where Theorem 4 shows NP-hardness even when $|S^+| = |S^-| = (k + 1)/2$, this does not hold in the discrete setting. We address this by constructing a reduction with an unbounded size for $S^+$.

**Theorem 6.** *The problem $k$-Counterfactual Explanation($\{0, 1\}, D_H$) is NP-complete for every odd integer $k \geqslant 1$.*

The NP upper bound is straightforward, so we focus on the hardness part of Theorem 6. We first establish the intractability of the following intermediate problem. Fix an integer $p \geqslant 0$. The input to the problem is given by a Boolean matrix $B$ of dimension $m \times n$ and an integer $\ell \leqslant n$. We want to know if it is possible to find a set $T \subseteq \{1, \ldots, n\}$ of size $|T| \leqslant \ell$ such that, after *flipping* every column from $B$ whose index is in $T$, the *weight* of at least $m - p$ 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 integer $p \geqslant 0$.*

*Proof.* Fix $p \geqslant 0$. We reduce from the following modified version of the *Vertex Cover* problem: Given an undirected graph $G = (V, E)$ and an integer $\ell$, is there a set $V' \subseteq V$ with $|V'| \leqslant \ell$ 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-complete 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 instance $(B, \ell + 1)$ for the problem $p$-BMCF.

Suppose first that $T \subseteq \{1, \ldots, |V| + 1\}$ is a solution for the instance $(B, \ell + 1)$ 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 $T' \subseteq \{1, \ldots, |V| + 1\}$ for $(B, \ell + 1)$ such that $|T| = |T'|$ and $T'$ contains index $|V| + 1$. Indeed, if $T$ does not contain index $|V| + 1$ then we can obtain such a solution $T'$ by simply removing any index from $T$ and adding index $|V| + 1$. Then $|T'| \leqslant \ell + 1$ and $|T''| \leqslant \ell$, 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

20

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 $|U| \leqslant \ell$ 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 $|T'| = |U| + 1 \leqslant \ell + 1$. We claim that $T'$ is a solution for the instance $(B, \ell + 1)$ 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 $|T'| - 1$. $\qquad \square$

*Proof of Theorem 6.* Fix an odd integer $k = 2p + 1$, for $p \geqslant 0$. 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 $\ell \leqslant n$. 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 $m \geqslant p + 1$. 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 $1 \leqslant j \leqslant p + 1$. 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 $(B, \ell)$ to $p$-BMCF has a solution $T \subseteq \{1, \ldots, n\}$. Then $|T| \leqslant \ell$. 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 $|T| \leqslant \ell$, 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 $d \leqslant \ell$ 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 $(B, \ell)$ 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. $\qquad \square$

**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 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 $X$ for $\bar{x}$ with respect to* $f^1_{S^+,S^-}$.

In contrast, $k$-CHECK SUFFICIENT REASON becomes intractable in this setting when $k \geq 3$.

**Theorem 7.** *The problem $k$-CHECK SUFFICIENT REASON$(\{0,1\}, D_H)$ is* coNP-*complete for every odd integer $k \geq 3$.*

*Proof.* The coNP upper bound is direct: given $S^+, S^-, \bar{x} \in \{0,1\}^n$ and $X \subseteq \{1, \ldots, n\}$, for all $\bar{y} \in \{0,1\}^n$ such that $\bar{y}[i] = \bar{x}[i]$, for every $i \in X$, we can verify in polynomial time whether $f^k_{S^+,S^-}(\bar{y}) = f^k_{S^+,S^-}(\bar{x})$.

We show that the complement of $k$-CHECK SUFFICIENT REASON$(\{0,1\}, D_H)$ is NP-hard for every fixed odd integer $k \geq 3$. We reduce from the Vertex Cover problem: given an undirected graph $G = (V, E)$ with $n$ vertices, and an integer $q \geq 0$, decide whether there is a subset of vertices $C \subseteq V$ with size $|C| \leq q$ such that $C$ is a vertex cover, that is, every edge in $E$ has an endpoint in $C$. We can assume without loss of generality that $q \geq n/2$. Indeed, if $q < n/2$, we consider the instance $(G', q' = n - q)$, where $G'$ results from $G$ by adding $n - 2q$ fresh nodes, each of them connected by an edge to every original node in $G$. Since any subset of vertices of size $n$ or $n - 1$ is a vertex cover, we can also assume that $q \leq n - 2$.

Take a graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and $E = \{e_1, \ldots, e_m\}$, for $m \geq 1$. The vector dimension is $n + (k+1)/2 + (2q - n)$ and then we write vectors in $\{0,1\}^{n+(k+1)/2+(2q-n)}$ as concatenations $(\bar{x}, \bar{y}, \bar{z})$ of three vectors $\bar{x} \in \{0,1\}^n$, $\bar{y} \in \{0,1\}^{(k+1)/2}$ and $\bar{z} \in \{0,1\}^{2q-n}$. 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 $S^-, S^+ \subseteq \{0,1\}^{n+(k+1)/2+(2q-n)}$ are defined as follows:

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

Note that $|S^+| = (k+1)/2$. Finally, we set $\bar{x} = (\bar{0}_n, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$. Note that $f^k_{S^+,S^-}(\bar{x}) = 0$ as $d_H(\bar{x}, (\bar{1}_n, \bar{\alpha}_h, \bar{0}_{2q-n})) = n + 1$ for every $h \in \{2, \ldots, (k+1)/2\}$, while

$$d_H(\bar{x}, (\bar{y}_1, \bar{\alpha}_h, \bar{1}_{2q-n})) = 2 + 1 + 2q - n \leq 3 + 2(n-2) - n = n - 1$$

for every $h \in \{1, \ldots, (k+1)/2\}$.

We claim that there is a vertex cover $C \subseteq \{1, \dots, n\}$ with $|C| \leqslant q$ if and only if the empty set is not a sufficient reason for $\bar{x}$ w.r.t $f_{S^+,S^-}^k$. The latter condition is equivalent to the existence of a vector $\bar{z} \in \{0,1\}^{n+(k+1)/2+(2q-n)}$ with $f_{S^+,S^-}^k(\bar{z}) = 1$. It suffices to show the following two properties:

1. If there is a vertex cover $C \subseteq \{1, \dots, n\}$ of $G$ with $|C| = q$, then $f_{S^+,S^-}^k(\bar{z}) = 1$, for $\bar{z} = (\bar{w}_C, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$, where $\bar{w}_C[i] = 0$ if $i \in C$, and $\bar{w}_C[i] = 1$ if $i \in \{1, \dots, n\} \setminus C$.

2. Suppose that $f_{S^+,S^-}^k(\bar{z}) = 1$ for $\bar{z} = (\bar{w}, \bar{\gamma}, \bar{t})$ and let $C = \{i \in \{1, \dots, n\} \mid \bar{w}[i] = 0\}$. Then we have only two possibilities: either $|C| \leqslant q$ and $C$ is a vertex cover of $G$; or $|C| = q+1$ and every subset $C' \subseteq C$ with $|C'| = q$ is a vertex cover of $G$.

These two properties imply the correctness of the reduction. If there is a vertex cover of size at most $q$, without loss of generality, we can assume the size is $q$ and then we can apply (1). For the backward direction, if $f_{S^+,S^-}^k(\bar{z}) = 1$ for some $\bar{z}$, property (2) implies that, in any of the possible two cases, there is a vertex cover of size at most $q$.

We start with property (1). Suppose $C$ is a vertex cover with $|C| = q$ and define $\bar{w}_C \in \{0,1\}^n$ such that $\bar{w}_C[i] = 0$ if $i \in C$ and $\bar{w}_C[i] = 1$, if $i \in \{1, \dots, n\} \setminus C$. We claim that $f_{S^+,S^-}^k(\bar{z}) = 1$ for the vector $\bar{z} = (\bar{w}_C, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$. For every $(\bar{y}_j, \bar{\beta}, \bar{1}_{2q-n}) \in S^-$, we have

$$d_H(\bar{z}, (\bar{y}_j, \bar{\beta}, \bar{1}_{2q-n})) \geqslant |V \setminus C| + 1 + 2q - n = n - q + 1 + 2q - n = q + 1.$$

Here we are using the fact that $C$ covers the edge $e_j$, that is, $|C \cap e_j| \geqslant 1$, and then:

$$d_H(\bar{w}_C, \bar{y}_j) = |C \cap e_j| + |V \setminus C| - (2 - |C \cap e_j|) \geqslant 1 + |V \setminus C| - (2 - 1) = |V \setminus C|.$$

On the other hand, we have:

$$d_H(\bar{z}, (\bar{0}_n, \bar{\alpha}_1, \bar{1}_{2q-n})) = |V \setminus C| + 1 + 2q - n = q + 1$$

and

$$d_H(\bar{z}, (\bar{1}_n, \bar{\alpha}_h, \bar{0}_{2q-n})) = |C| + 1 = q + 1$$

for every $h \in \{2, \dots, (k+1)/2\}$.

We now turn to property (2). Suppose that $f_{S^+,S^-}^k(\bar{z}) = 1$, for some $\bar{z} = (\bar{w}, \bar{\gamma}, \bar{t}) \in \{0,1\}^{n+(k+1)/2+(2q-n)}$ and define $C = \{i \in \{1, \dots, n\} \mid \bar{w}[i] = 0\}$. We can always pick vectors $\bar{\beta}_1, \dots, \bar{\beta}_{(k+1)/2}$ from $\{0,1\}^{(k+1)/2} \setminus \{\bar{0}_{(k+1)/2}\}$ such that $d_H(\bar{\gamma}, \bar{\beta}_h) \leqslant 1$, for every $h \in \{1, \dots, (k+1)/2\}$.

By contradiction, suppose first that $|C| \geqslant q + 2$. Then:

$$\begin{aligned} d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{y}_1, \bar{\beta}_h, \bar{1}_{2q-n})) &\leqslant d_H(\bar{w}, \bar{y}_1) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\ &= |C \cap e_1| + |V \setminus C| - (2 - |C \cap e_1|) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\ &\leqslant 2 + |V \setminus C| + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\ &\leqslant 2 + (n - q - 2) + 1 + 2q - n \\ &= q + 1 \end{aligned}$$

for every $h \in \{1, \dots, (k+1)/2\}$. On the other hand, we have

$$d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{1}_n, \bar{\alpha}_h, \bar{0}_{2q-n})) \geqslant d_H(\bar{w}, \bar{1}_n) = |C| \geqslant q + 2$$

for every $h \in \{2, \dots, (k+1)/2\}$. By Proposition 1, choosing $A = \{(\bar{y}_1, \bar{\beta}_h, \bar{1}_{2q-n}) \mid h \in \{1, \dots, (k+1)/2\}\} \subseteq S^-$ and $B = \{(\bar{0}_n, \bar{\alpha}_1, \bar{1}_{2q-n})\} \subseteq S^+$, we have that $f_{S^+,S^-}^k(\bar{z}) = 0$; a contradiction.

Assume that $|C| \leqslant q$ and, towards a contradiction, suppose that $C$ is not a vertex cover. There must be an edge $e_j$ such that $|C \cap e_j| = 0$. We have that

$$
\begin{aligned}
d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{y}_j, \bar{\beta}_h, \bar{1}_{2q-n})) &\leqslant d_H(\bar{w}, \bar{y}_j) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\
&= |C \cap e_j| + |V \setminus C| - (2 - |C \cap e_j|) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\
&= |V \setminus C| - 1 + d_H(\bar{t}, \bar{1}_{2q-n})
\end{aligned}
$$

for every $h \in \{1, \dots, (k+1)/2\}$. On the other hand, we have

$$
d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{0}_n, \bar{\alpha}_1, \bar{1}_{2q-n})) \geqslant d_H(\bar{w}, \bar{0}_n) + d_H(\bar{t}, \bar{1}_{2q-n}) = |V \setminus C| + d_H(\bar{t}, \bar{1}_{2q-n}).
$$

We conclude that $f_{S^+, S^-}^k(\bar{z}) = 0$, using Proposition 1 with $A = \{(\bar{y}_j, \bar{\beta}_h, \bar{1}_{2q-n}) \mid h \in \{1, \dots, (k+1)/2\}\} \subseteq S^-$ and $B = S^+ \setminus \{(\bar{0}_n, \bar{\alpha}_1, \bar{1}_{2q-n})\}$. This is a contradiction.

Finally, suppose that $|C| = q + 1$. We claim that there is no edge with an endpoint in $V \setminus C$. By contradiction, suppose there is such an edge, say $e_j$. In particular, $|C \cap e_j| \leqslant 1$. We obtain:

$$
\begin{aligned}
d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{y}_j, \bar{\beta}_h, \bar{1}_{2q-n})) &\leqslant d_H(\bar{w}, \bar{y}_j) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\
&= |C \cap e_j| + |V \setminus C| - (2 - |C \cap e_j|) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\
&\leqslant 1 + |V \setminus C| - (2 - 1) + 1 + d_H(\bar{t}, \bar{1}_{2q-n}) \\
&\leqslant (n - q - 1) + 1 + 2q - n \\
&= q
\end{aligned}
$$

for every $h \in \{1, \dots, (k+1)/2\}$. On the other hand, we have

$$
d_H((\bar{w}, \bar{\gamma}, \bar{t}), (\bar{1}_n, \bar{\alpha}_h, \bar{0}_{2q-n})) \geqslant d_H(\bar{w}, \bar{1}_n) = |C| = q + 1
$$

for every $h \in \{2, \dots, (k+1)/2\}$. Again, by Proposition 1, we conclude that $f_{S^+, S^-}^k(\bar{z}) = 0$; a contradiction. It follows that every edge of $G$ is contained in $C$, and, since $|C| = q + 1$, any subset $C' \subseteq C$ with $|C'| = q$ is a vertex cover of $G$. $\qquad\square$

The hardness of $k$-Check Sufficient Reason$(\{0, 1\}, D_H)$ holds even when the input subset of components $X \subseteq \{1, \dots, n\}$ is the empty set. As a consequence, $k$-Minimal Sufficient Reason$(\{0, 1\}, D_H)$ is also hard when $k \geqslant 3$. This also implies hardness for the problem of computing minimal sufficient reasons.

**Corollary 5.** *The problem $k$-Minimal Sufficient Reason$(\{0, 1\}, D_H)$ is* coNP-*hard for every odd integer $k \geqslant 3$.*

# 8  Further Complexity Results

We start by noting that some of the NP-hardness results presented in Section 4 for minimum sufficient reasons are actually completeness results. Indeed, the results in Sections 5, 6 and 7 show that checking sufficient reasons can be done in polynomial time in several settings. This directly gives us NP algorithms for these cases.

**Corollary 6.** *The following problems are* NP-*complete:* 1-Minimum Sufficient Reason$(\{0, 1\}, D_H)$, 1-Minimum Sufficient Reason$(\mathbb{R}, D_1)$, *and* $k$-Minimum Sufficient Reason$(\mathbb{R}, D_2)$, *for every fixed odd integer $k \geqslant 1$.*

In the remaining of the section we establish two additional complexity results. We first pinpoint the complexity of $k$-Minimum Sufficient Reason$(\{0, 1\}, D_H)$, when $k \geqslant 3$, and show that this problem is complete for $\Sigma_2^p$, the second level of the polynomial hierarchy. Then we study the parameterized complexity of $k$-Counterfactual Explanation in the setting based on $\ell_2$-distance.

**The complexity of Minimum Sufficient Reason in the discrete setting, for $k \geqslant 3$**   We assume familiarity with the complexity class $\Sigma_2^p$, which corresponds to $\mathrm{NP}^{\mathrm{NP}}$, that is, $\mathrm{NP}$ with an oracle in $\mathrm{NP}$. We show the following:

**Theorem 8.** $k$-Minimum Sufficient Reason$(\{0,1\}, D_H)$ *is $\Sigma_2^p$-complete, for every fixed odd integer $k \geqslant 3$.*

We start by establishing the $\Sigma_2^p$-completeness of the following auxiliary problem:

| | |
|---|---|
| PROBLEM : | $\exists \not\forall$-Vertex-Cover |
| INPUT : | Graph $G = (V, E)$, and two positive integers $p < q$. |
| OUTPUT : | Yes, if there is $S \subseteq V$, with $|S| \leqslant p$, such that no superset $S'$ ($S \subseteq S' \subseteq V$) with $|S'| \leqslant q$ is a vertex cover of $G$. No otherwise. |

This problem is tightly related to *"interdiction"* optimization problems over graphs, in which one has some base problem (i.e., min vertex-cover, max clique, min dominating set, etc.) and the task is to find a small subset of the vertices that intersects every optimal solution. A nice general framework for proving $\Sigma_2^p$-completeness of interdiction problems has been recently established by Grüne and Wulf [Grüne and Wulf(2025)]. For our purposes, nonetheless, it will be enough to consider a simpler and older result from [Rutenburg(1994)]: the independent set interdiction problem, defined next, is $\Sigma_2^p$-complete.

| | |
|---|---|
| PROBLEM : | Independent Set Interdiction |
| INPUT : | Graph $G = (V, E)$, and two positive integers $p, q$. |
| OUTPUT : | Yes, if there is $S \subseteq V$, with $|S| \leqslant p$, such that every independent set $S' \subseteq V$ with $|S'| \geqslant q$ holds $S \cap S' \neq \varnothing$. No otherwise. |

**Theorem 9.** *The $\exists \not\forall$-Vertex-Cover problem is $\Sigma_2^p$-complete. Hardness holds even when $n/2 \leqslant q \leqslant n - 2$, where $n$ is the number of vertices of the input graph.*

*Proof.* Membership is direct: the NP-part of the computation guesses $S$, checks its size, and calls a coNP-oracle to check that every superset $S' \supseteq S$ of size at most $q$ fails to cover $G$.

For hardness, we reduce from Independent Set Interdiction. Let $(G, p, q)$ be an input instance, where $G = (V, E)$. We then simply output $(G, p, |V| - q)$, which obviously can be done in polynomial time.

To prove correctness we will need some notation. Given a graph $G$, we denote by $\alpha(G)$ the maximum size of an independent set of $G$, and by $\tau(G)$ the minimum size of a vertex cover of $G$. Moreover, given a graph $G$ and a subset $S \subseteq V$, we denote by $\tau(G, S)$ the minimum size of a vertex cover that contains $S$. Our reduction relies on two simple observations:

1. $\alpha(G) + \tau(G) = |V|$.

2. $\tau(G, S) = |S| + \tau(G[V \setminus S])$.

While observation (1) corresponds to the well-known fact that the complement of a vertex cover is an independent set (and vice versa), observation (2) requires a brief justification. Indeed, note that if $S' \supseteq S$ is any vertex cover of size $\tau(G, S)$, then $S' \setminus S$ must be a vertex cover of $G[V \setminus S]$, and thus $\tau(G, S) = |S'| = |S| + |S' \setminus S| \geqslant |S| + \tau(G[V \setminus S])$. Conversely, let $S''$ be a vertex cover of minimum size for $G[V \setminus S]$. Then, $S'' \cup S$ is a vertex cover for $G$ which contains $S$, and thus $|S'' \cup S| = \tau(G[V \setminus S]) + |S| \geqslant \tau(G, S)$.

Now, observe that $\exists \not\forall$-Vertex-Cover can be rephrased as whether there exists $S \subseteq V$ with $|S| \leqslant p$ such that $\tau(G, S) > q$. Similarly, the Independent Set Interdiction problem can be rephrased as whether there

exists $S \subseteq V$ with $|S| \leqslant p$ such that $\alpha(G[V \setminus S]) < q$, since the independent sets of $G$ that do not intersect $S$ are exactly the independent sets of $G[V \setminus S]$. Using these rephrasings to prove the correctness of our reduction, it suffices to show that for any $S \subseteq V$ with $|S| \leqslant p$, we have

$$\tau(G, S) > q \iff \alpha(G[V \setminus S]) < |V| - q.$$

This now follows easily from our observations:

$$
\begin{aligned}
\tau(G, S) > q &\iff |S| + \tau(G[V \setminus S]) > q && \text{(Observation (2))}\\
&\iff |S| + (|V \setminus S| - \alpha(G[V \setminus S])) > q && \text{(Observation (1))}\\
&\iff |S| + (|V| - |S| - \alpha(G[V \setminus S])) > q \\
&\iff \alpha(G[V \setminus S]) < |V| - q.
\end{aligned}
$$

It remains to show that hardness holds when $n/2 \leqslant q \leqslant n - 2$. Let $(G, p, q)$ be an instance of $\exists\nvdash$-Vertex-Cover, where $G = (V, E)$ and $n := |V|$. If $q < n/2$, we transform the instance to $(G', p, q')$, where $q' = n - q$ and $G' = (V', E')$ results from $G$ by adding $n - 2q$ fresh nodes, each of them connected by an edge to every node in $G$. Note that $q' = n'/2$ where $n' = |V'|$. We claim that $(G, p, q) \in \exists\nvdash$-Vertex-Cover if and only if $(G', p, q') \in \exists\nvdash$-Vertex-Cover. Suppose first that there is $S \subseteq V$ with $|S| \leqslant p$ such that every superset $S'$ of $S$ with $|S'| \leqslant q$ is not a vertex cover of $G$. It holds that there is no superset $S''$ of $S$ with $|S''| \leqslant q'$ that is a vertex cover of $G'$. By contradiction, if there is such a $S''$, as $q' < n$ and there is an edge from every fresh node to every node in $V$, then $S''$ must contain all the $n - 2q$ fresh nodes. It follows that $|S'' \cap V| \leqslant q$. This is a contradiction as $S'' \cap V$ is a vertex cover of $G$ and a superset of $S$. On the other hand, assume there is $S \subseteq V'$ with $|S| \leqslant p$ such that every superset $S'$ of $S$ with $|S'| \leqslant q'$ is not a vertex cover of $G'$. We claim that there is no superset $S'' \subseteq V$ of $S \cap V$ with $|S''| \leqslant q$ that is a vertex cover of $G$. If there is such a $S''$ then by adding the fresh nodes, we obtain a superset of $S$ that is a vertex cover of $G'$ of size at most $q + n - 2q = q'$; a contradiction. Finally, note that any instance $(G, p, q)$ with $q \geqslant n - 1$ is a negative instance of $\exists\nvdash$-Vertex-Cover, as any set of at least $n - 1$ nodes is a vertex cover of $G$. $\qquad\square$

*Proof of Theorem 8.* The upper bound is direct: we guess a subset $X$ of components with $|X| \leqslant \ell$ and then verify with an coNP-oracle that is a sufficient reason (recall Theorem 7).

We reduce from $\exists\nvdash$-Vertex-Cover under the restriction $n/2 \leqslant q \leqslant n - 2$. Given an instance $(G, p, q)$ of $\exists\nvdash$-Vertex-Cover, where $G = (V, E)$ and $V = \{1, \ldots, n\}$, we define the instance $(S^+, S^-, \bar{x}, p)$ of $k$-Minimum Sufficient Reason$(\{0, 1\}, D_H)$, where $(S^+, S^-, \bar{x})$ is define from $(G, q)$ *exactly* as in the proof of Theorem 7. Recall that the dimension of vectors in $(S^+, S^-, \bar{x})$ is $n + (k + 1)/2 + (2q - n)$, and hence we write vectors as the concatenation of three vectors of dimensions $n$, $(k + 1)/2$ and $(2q - n)$, respectively. The vector $\bar{x}$ is the all zero vector $\bar{x} = (\bar{0}_n, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$ (here $\bar{0}_r$ is the all zero vector of dimension $r$). It holds that $f^k_{S^+, S^-}(\bar{x}) = 0$. The condition $n/2 \leqslant q \leqslant n - 2$ is needed to apply this reduction. The following properties of $(S^+, S^-, \bar{x})$ are shown in the proof of Theorem 7:

1. If there is a vertex cover $C \subseteq \{1, \ldots, n\}$ of $G$ with $|C| = q$, then $f^k_{S^+, S^-}(\bar{z}) = 1$, for $\bar{z} = (\bar{w}_C, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$, where $\bar{w}_C[i] = 0$ if $i \in C$, and $\bar{w}_C[i] = 1$ if $i \in \{1, \ldots, n\} \setminus C$.

2. Suppose that $f^k_{S^+, S^-}(\bar{z}) = 1$ for $\bar{z} = (\bar{w}, \bar{\gamma}, \bar{t})$ and let $C = \{i \in \{1, \ldots, n\} \mid \bar{w}[i] = 0\}$. Then, either $|C| \leqslant q$ and $C$ is a vertex cover of $G$; or $|C| = q + 1$ and every subset $C' \subseteq C$ with $|C'| = q$ is a vertex cover of $G$.

To see that the reduction is correct, assume first that there exists $S \subseteq \{1, \ldots, n\}$ with $|S| \leqslant p$ such that no superset $S'$ of $S$ with $|S'| \leqslant q$ is a vertex cover of $G$. We claim that $S$ is a sufficient reason for $\bar{x}$ w.r.t. $f^k_{S^+, S^-}$. By contradiction, suppose there is a vector $\bar{z} = (\bar{w}, \bar{\gamma}, \bar{t})$ with $\bar{z}[i] = \bar{x}[i] = 0$, for every $i \in S$, such that $f^k_{S^+, S^-}(\bar{z}) = 1$. Define $C = \{i \in \{1, \ldots, n\} \mid \bar{w}[i] = 0\}$. It holds that $S \subseteq C$. By property (2), we have two possible cases. If $|C| \leqslant q$, then $C$ is a vertex cover, and we obtain a contradiction. If $|C| = q + 1$, we pick any $u \in C \setminus S$ (recall $p < q$), and then $C \setminus \{u\}$ is a superset of $S$ with $|C \setminus \{u\}| = q$ that is a vertex cover; a contradiction.

For the other direction, suppose that $X \subseteq \{1, \ldots, n + (k+1)/2 + (2q - n)\}$ is a sufficient reason for $\bar{x}$ with $|X| \leqslant p$. Take $S = X \cap \{1, \ldots, n\}$. We claim that no superset $C$ of $S$ with $|C| \leqslant q$ is a vertex cover of $G$. By contradiction, suppose that such a $C$ exists. We can assume that $|C| = q$. Take the vector $\bar{z} = (\bar{w}_C, \bar{0}_{(k+1)/2}, \bar{0}_{2q-n})$, where $\bar{w}_C[i] = 0$, if $i \in C$, and $\bar{w}_C[i] = 1$, otherwise. It holds that $\bar{z}[i] = \bar{x}[i] = 0$, for every $i \in X$. Moreover, by property (1), we obtain that $f^k_{S^+,S^-}(\bar{z}) = 1$. This is a contradiction. □

# 9 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.

## 9.1 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` [Gurobi Optimization, LLC(2024)] solver for IQP, and the recent `cardinality-cadical` solver [Reeves et al.(2024)] 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 [Deng(2012)] and on synthetic random data. For the MNIST dataset, we consider both the original grayscale $28 \times 28$ 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.

## 9.2 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 [Douze et al.(2024)]. 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. [Contardo et al.(2024)]. For counterfactual explanations in the discrete setting, we consider first the following IQP formulation for a vector $\bar{x}$ classified positively:

$$\text{minimize} \quad \sum_{i=1}^{n} (\bar{x}[i] - \bar{y}[i])^2$$

$$\text{subject to } d^+ = \min_{\bar{z} \in S^+} \sum_{i=1}^{n} (\bar{z}[i] - \bar{y}[i])^2,$$

$$d^- = \min_{\bar{z} \in S^-} \sum_{i=1}^{n} (\bar{z}[i] - \bar{y}[i])^2,$$

$$d^- < d^+,$$

$$\bar{y} \in \{0, 1\}^d, \ d^+ \in \mathbb{R}, \ d^- \in \mathbb{R},$$

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 $m \leqslant r_i$ and $v_i \cdot r_i \leqslant m$, 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 [Reeves et al.(2024)]. Given boolean literals $\ell_1, \ldots, \ell_n$, cardinality constraints are of the form $\sum_{i=1}^{n} \ell_i \geqslant b$, for some constant integer $b$ (the "bound"). On the other hand, "guarded" cardinality constraints, provided a different boolean literal $g$, are of the form $g \implies \left( \sum_{i=1}^{n} \ell_i \geqslant b \right)$. 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

$$c_i \implies d_H(\bar{y}, \bar{o}) < d_H(\bar{y}, \bar{s}), \quad \forall \bar{s} \in S^+, \tag{8}$$

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 (8) as the following guarded cardinality constraint:

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

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

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

is equivalent to $d_H(\bar{x}, \bar{y}) \leqslant k$. 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 5 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 6 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 [Douze et al.(2024)] 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 [Contardo et al.(2024), Eppstein(2022)], is part of our future work.
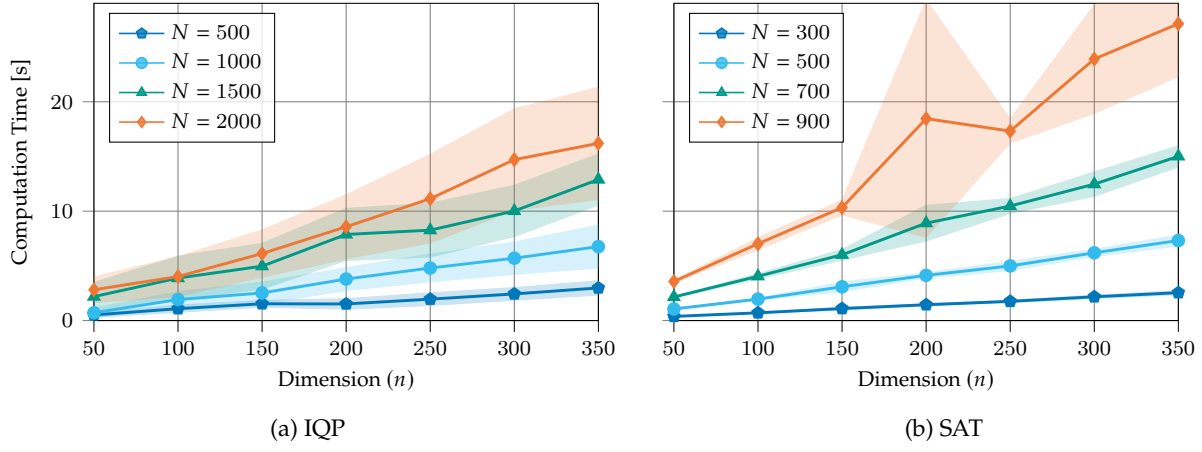
(a) IQP

(b) SAT

Figure 5: 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 ($\ell_1$)
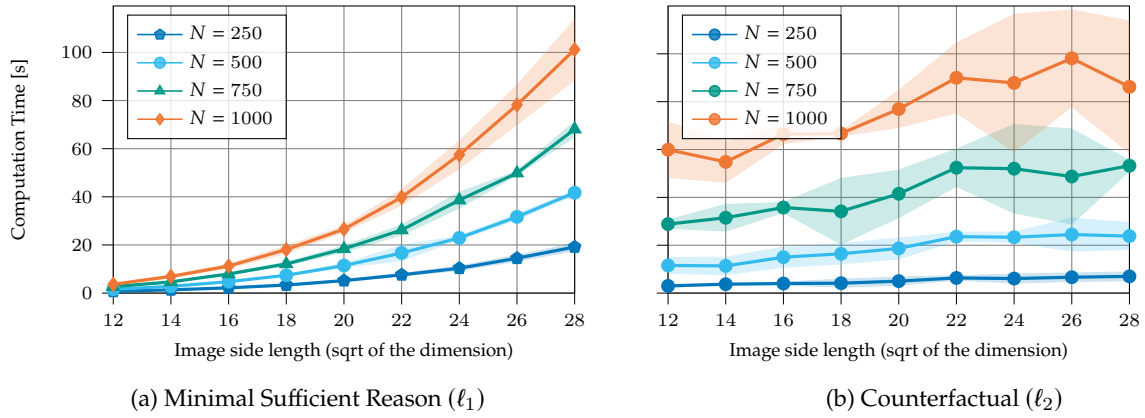
(b) Counterfactual ($\ell_2$)

Figure 6: 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.

| Explanation | Counterfactual | Check Sufficient Reason | | Minimum Sufficient Reason | |
|---|---|---|---|---|---|
| **Metric space** | $k \geqslant 1$ | $k = 1$ | $k > 1$ | $k = 1$ | $k > 1$ |
| $(\mathbb{R}, D_2)$ | P (Thm. 2) | P (Prop. 3) | P (Prop. 3) | NP-c (Coro. 6) | NP-c (Coro. 6) |
| $(\mathbb{R}, D_1)$ | NP-c (Thm. 4) | P (Prop. 4) | coNP-c (Thm. 5) | NP-c (Coro. 6) | NP-h (Thm. 1) |
| $(\{0,1\}, D_H)$ | NP-c (Thm. 6) | P (Prop. 6) | coNP-c (Thm. 7) | NP-c (Coro. 6) | $\Sigma_2^p$-c (Thm. 8) |

Table 1: Summary of complexity results.

## 10   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 problems of checking and computing minimal sufficient reasons are polynomial-time solvable when checking sufficient reasons is tractable (recall Proposition 2), and hard otherwise. All results in Table 1 are completeness results, except for minimum sufficient reasons under the $\ell_1$-distance and $k > 1$, where we only obtain NP-hardness. We leave the precise complexity of this case open.

The kind of explanations we have studied are often said to be "local" [Guidotti et al.(2018), Hakkoum et al.(2024)], 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 [Bassan et al.(2024)]. Interestingly, a recent line of work has studied the computational problem of *thinning* $k$-NN classifiers by removing redundant points in the training set [Eppstein(2022), Flores-Velazco(2022), Rohrer and Weber(2023)]. 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 seek to explore $k$-COUNTERFACTUAL EXPLANATION for metrics based on $\ell_p$, where $p > 2$. Specifically, we ask whether $\ell_2$ is the only metric for which this problem is tractable.

- Second, it is interesting to explore, whether our positive results are extendable to the case of more than two labels. It does not make any difference for $k = 1$: if a point $\bar{x}$ to be explained was classified with some label, then all the other labels can be merged into a single one, effectively reducing everything to the binary case. The same reasoning does not work for $k \geqslant 3$, and it is open, whether the $k$-COUNTERFACTUAL EXPLANATION can still be computed in polynomial time for $\ell_2$ in the multi-label case. One can generalize our algorithm from Theorem 2 to have exponential dependence on the number of labels; this answers our question positively when the number of labels is constant, but not in the general case.

- 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?

# References

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

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

[Agarwal et al.(2016)] Pankaj K. Agarwal, Kyle Fox, Kamesh Munagala, and Abhinandan Nath. 2016. Parallel Algorithms for Constructing Range and Nearest-Neighbor Searching Data Structures. In *PODS*. ACM, 429–440.

[Arenas et al.(2022)] Marcelo Arenas, Pablo Barceló, Miguel A. Romero Orth, and Bernardo Subercaseaux. 2022. On Computing Probabilistic Explanations for Decision Trees. In *NeurIPS*.

[Arrieta et al.(2020)] 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. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[Artelt and Hammer(2019)] André Artelt and Barbara Hammer. 2019. On the computation of counterfactual explanations–A survey. *arXiv preprint arXiv:1911.07749* (2019).

[Audemard et al.(2022)] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. 2022. On the Explanatory Power of Boolean Decision Trees. *Data & Knowledge Engineering* 142 (Nov. 2022), 102088. doi:10.1016/j.datak.2022.102088

[Audemard et al.(2024)] Gilles Audemard, Jean-Marie Lagniez, and Pierre Marquis. 2024. On the Computation of Contrastive Explanations for Boosted Regression Trees. In *ECAI*. 1083–1091.

[Barceló et al.(2020)] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model Interpretability through the lens of Computational Complexity. In *NeurIPS*. 15487–15498.

[Bassan et al.(2024)] Shahaf Bassan, Guy Amir, and Guy Katz. 2024. Local vs. global interpretability: a computational complexity perspective. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 126, 35 pages.

[Beyer et al.(1999)] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When Is "Nearest Neighbor" Meaningful?. In *ICDT*. 217–235.

[Biau and Devroye(2015)] Gérard Biau and Luc Devroye. 2015. *Lectures on the Nearest Neighbor Method*. Springer International Publishing. doi:10.1007/978-3-319-25388-6

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

[Contardo et al.(2024)] Claudio Contardo, Ricardo Fukasawa, Louis-Martin Rousseau, and Thibaut Vidal. 2024. Optimal Counterfactual Explanations for k-Nearest Neighbors Using Mathematical Optimization and Constraint Programming. In *ISCO*. 318–331.

[Cygan et al.(2015)] Marek Cygan, Fedor V Fomin, Łukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk, and Saket Saurabh. 2015. *Parameterized algorithms*. Vol. 5. Springer.

[Darwiche(2023)] Adnan Darwiche. 2023. Logic for Explainable AI. In *LICS*. IEEE, 1–11.

[Darwiche and Hirth(2020)] Adnan Darwiche and Auguste Hirth. 2020. On The Reasons Behind Decisions. *CoRR* abs/2002.09284 (2020). https://arxiv.org/abs/2002.09284

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

[Douze et al.(2024)] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[Eppstein(2022)] David Eppstein. 2022. Finding Relevant Points for Nearest-Neighbor Classification. In *2022 Symposium on Simplicity in Algorithms (SOSA)*. Society for Industrial and Applied Mathematics, 68–78. doi:10.1137/1.9781611977066.6

[Fan and Koutris(2022)] Austen Z. Fan and Paraschos Koutris. 2022. Certifiable Robustness for Nearest Neighbor Classifiers. In *ICDT (LIPIcs)*. 6:1–6:20.

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

[Forel et al.(2023)] Alexandre Forel, Axel Parmentier, and Thibaut Vidal. 2023. Explainable Data-Driven Optimization: From Context to Decision and Back Again. In *ICML*, Vol. 202. PMLR, 10170–10187.

[Garey and Johnson(1979)] Michael R Garey and David S Johnson. 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness.

[Grüne and Wulf(2025)] Christoph Grüne and Lasse Wulf. 2025. The Complexity of Blocking All Solutions. arXiv:2502.05348 [cs.CC] https://arxiv.org/abs/2502.05348

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

[Guidotti et al.(2018)] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. doi:10.1145/3236009

[Gurobi Optimization, LLC(2024)] Gurobi Optimization, LLC. 2024. Gurobi Optimizer Reference Manual. https://www.gurobi.com

[Hakkoum et al.(2024)] Hajar Hakkoum, Ali Idri, and Ibtissam Abnane. 2024. Global and Local Interpretability Techniques of Supervised Machine Learning Black Box Models for Numerical Medical Data. *Engineering Applications of Artificial Intelligence* 131 (May 2024), 107829. doi:10.1016/j.engappai.2023.107829

[Han et al.(2023)] Yunqiu Han, Yizuo Chen, and Adnan Darwiche. 2023. On the Complexity of Counterfactual Reasoning. In *IJCAI*. ijcai.org, 5676–5684.

[Hinneburg et al.(2000)] Alexander Hinneburg, Charu C. Aggarwal, and Daniel A. Keim. 2000. What Is the Nearest Neighbor in High Dimensional Spaces?. In *VLDB*. 506–515.

[Hiriart-Urruty and Lemaréchal(2001)] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. 2001. *Fundamentals of Convex Analysis*. Springer, Berlin, Heidelberg.

[Huang et al.(2021)] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and João Marques-Silva. 2021. On Efficiently Explaining Graph-Based Classifiers. In *KR*. 356–367.

[Ignatiev and Marques-Silva(2021)] Alexey Ignatiev and Joao Marques-Silva. 2021. SAT-Based Rigorous Explanations for Decision Lists. In *Theory and Applications of Satisfiability Testing – SAT 2021*. Springer International Publishing, 251–269. doi:10.1007/978-3-030-80223-3_18

[Ignatiev et al.(2021)] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. 2021. From Contrastive to Abductive Explanations and Back Again. In *AIxIA 2020 – Advances in Artificial Intelligence*. Springer International Publishing, 335–355. doi:10.1007/978-3-030-77091-4_21

[Ignatiev et al.(2019)] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. 2019. Abduction-Based Explanations for Machine Learning Models. In *AAAI*. 1511–1519.

[Izza et al.(2023)] Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin Cooper, and Joao Marques-Silva. 2023. On Computing Probabilistic Abductive Explanations. *International Journal of Approximate Reasoning* 159 (2023), 108939. doi:10.1016/j.ijar.2023.108939

[Izza et al.(2022a)] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. 2022a. On Tackling Explanation Redundancy in Decision Trees. *Journal of Artificial Intelligence Research* 75 (Sept. 2022), 261–321. doi:10.1613/jair.1.13575

[Izza et al.(2021)] Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and J. Marques-Silva. 2021. Efficient Explanations With Relevant Sets. *ArXiv* abs/2106.00546 (2021).

[Izza et al.(2022b)] Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. 2022b. Provably Precise, Succinct and Efficient Explanations for Decision Trees. doi:10.48550/ARXIV.2205.09569

[Jing et al.(2024)] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. 2024. When Large Language Models Meet Vector Databases: A Survey. *CoRR* abs/2402.01763 (2024).

[Kozlov et al.(1980)] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. 1980. The polynomial solvability of convex quadratic programming. *U. S. S. R. Comput. Math. and Math. Phys.* 20, 5 (Jan. 1980), 223–228. doi:10.1016/0041-5553(80)90098-1

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

[Lewis and Papadimitriou(1997)] Harry R Lewis and Christos H Papadimitriou. 1997. Elements of the Theory of Computation.

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

[Magagnini et al.(2025)] Marica Magagnini, Emilio Carrizosa, and Renato De Leone. 2025. Nearest Neighbors Counterfactuals. In *Machine Learning, Optimization, and Data Science*, Giuseppe Nicosia, Varun Ojha, Sven Giesselbach, M. Panos Pardalos, and Renato Umeton (Eds.). Springer Nature Switzerland, Cham, 193–208. doi:10.1007/978-3-031-82481-4_14

[Marques-Silva et al.(2020)] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *NeurIPS*.

[Marques-Silva et al.(2021)] Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, and Nina Narodytska. 2021. Explanations for Monotonic Classifiers.. In *ICML*. 7469–7479.

[Marques-Silva et al.(2021)] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. 2021. Explanations for Monotonic Classifiers. In *ICML*. 7469–7479.

[Marques-Silva and Ignatiev(2022)] Joao Marques-Silva and Alexey Ignatiev. 2022. Delivering Trustworthy AI through Formal XAI. In *AAAI*.

[Molnar(2022)] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book

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

[Ribeiro et al.(2018)] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*. 1527–1535.

[Rohrer and Weber(2023)] Josiah Rohrer and Simon Weber. 2023. Reducing Nearest Neighbor Training Sets Optimally and Exactly. *arXiv preprint arXiv:2302.02132* (2023).

[Roussopoulos et al.(1995)] Nick Roussopoulos, Stephen Kelley, and Frédéic Vincent. 1995. Nearest Neighbor Queries. In *SIGMOD*. ACM Press, 71–79.

[Rutenburg(1994)] Vladislav Rutenburg. 1994. Propositional Truth Maintenance Systems: Classification and Complexity Analysis. *Annals of Mathematics and Artificial Intelligence* 10, 3 (Sept. 1994), 207–231. doi:10.1007/BF01530952

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

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

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

[Wäldchen et al.(2021)] Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. 2021. The Computational Complexity of Understanding Binary Classifier Decisions. *J. Artif. Intell. Res.* 70 (2021), 351–387.