

Why this and not that? A Logic-based Framework for Contrastive Explanations

Tobias Geibinger¹[0000–0002–0856–7162], Reijo Jaakkola²[0000–0003–4714–4637],
Antti Kuusisto²[0000–0003–1356–8749], Xinghan Liu¹[0000–0002–5533–3924], and
Miikka Vilander²[0000–0002–7301–939X]

¹ TU Wien, Austria

{tobias.geibinger,xinghan.liu}@tuwien.ac.at

² Tampere University, Finland

{reijo.jaakkola,antti.kuusisto}@tuni.fi

Abstract. We define several canonical problems related to contrastive explanations, each answering a question of the form “Why P but not Q ?”. The problems compute causes for both P and Q , explicitly comparing their differences. We investigate the basic properties of our definitions in the setting of propositional logic. We show, inter alia, that our framework captures a cardinality-minimal version of existing contrastive explanations in the literature. Furthermore, we provide an extensive analysis of the computational complexities of the problems. We also implement the problems for CNF-formulas using answer set programming and present several examples demonstrating how they work in practice.

Keywords: Explainable AI · Contrastive Explanations · Logic · Answer Set Programming

1 Introduction

The importance of explanations for decisions made by automatic classifiers has been well-established with the rise of AI methods. In this work, we investigate a category of explanations called *contrastive explanations*, which answer the question “Why P , but not Q ?” These types of questions are very common in practical contexts, when an expected outcome was not obtained. It has also been argued [14] that even when not explicitly asking for one, people often prefer an explanation in the form of a comparison between the situation as it occurred in reality and a different one that could have happened.

In this work, we use a logic-based framework to formalize several problems where the task is to find contrastive explanations. Related problems have been studied previously, for example, by Darwiche [4] and Ignatiev et al. [9], where, broadly speaking, the goal is to find a *counterfactual explanation*: “Had H been true, it would have been the case that Q ”. In contrast to these works, solutions to our problems explicitly answer both “Why P ?” and “Why not Q ?” with dedicated output formulas that are required to be structurally similar. Our definitions have been partially inspired by the work of Lipton [12], where he argues that a

contrastive explanation should contrast the causes of P against the absence of corresponding causes of Q .

Our first problem, which we call the contrastive explanation problem, aims to explain why two seemingly similar entities have differing properties. For example, we might ask why two individuals with similar backgrounds were assigned different credit risk levels. An input to our problem consists of two sets of formulas S, S' along with two formulas φ, ψ such that $S \models \varphi \wedge \neg\psi$ and $S' \models \neg\varphi \wedge \psi$. The output of the problem is a minimal size triple (θ, θ', χ) of formulas, such that $\theta \wedge \chi$ explains why S implies $\varphi \wedge \neg\psi$, contrasting with $\theta' \wedge \chi$ that explains why S' implies $\neg\varphi \wedge \psi$. Formally, we require that $S \models \theta \wedge \chi \models \varphi \wedge \neg\psi$ and $S' \models \theta' \wedge \chi \models \neg\varphi \wedge \psi$. Our problem formulation naturally enforces the similarity of the two explanations, as it encourages shifting content from θ and θ' into the common formula χ . To see the intuition behind this, consider the question why A is a dog and B is a cat. Here one would not insert “cannot fly” into the differentiating formulae θ and θ' , since neither A nor B can fly. On the other hand, it might be necessary to include “cannot fly” to χ , since it separates cats and dogs from, say, crows.

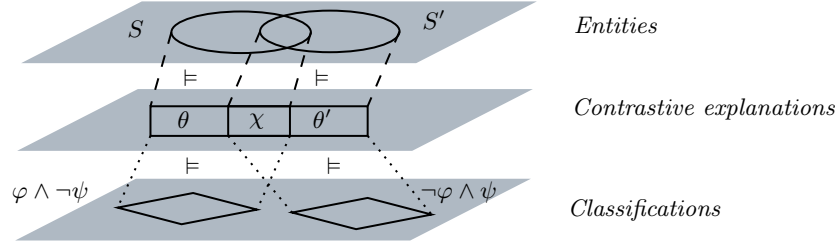


Fig. 1. Illustration of the contrastive explanation problem

Setting $S = \{\varphi \wedge \neg\psi\}$ and $S' = \{\neg\varphi \wedge \psi\}$ in the contrastive explanation problem, we obtain as a special case a problem that directly compares all the models of $\varphi \wedge \neg\psi$ with those of $\neg\varphi \wedge \psi$, giving a global contrast between the two formulas. Accordingly, we call this the global contrastive problem. We also define an alternative global problem which we call the minimal separator problem. In this problem the goal is to find a single minimal difference that suffices to separate the two input formulas.

We then move on to consider a variant of this problem where the input does not contain S' and the question we want to answer is “why does S satisfy φ but not ψ ?”. We formalize this problem in two distinct ways, both rooted in counterfactual reasoning. In the first problem, which we call the counterfactual contrastive explanation problem, the output is a minimal size triple (θ, θ', χ) such that $S \models \theta \wedge \chi \models \varphi \wedge \neg\psi$ and $\theta' \wedge \chi \models \neg\varphi \wedge \psi$. The second problem, called the counterfactual difference problem, is otherwise the same except that we require that $S \equiv \theta \wedge \chi$, i.e., that $\theta \wedge \chi$ defines the current state of affairs S . Roughly

speaking, in the first problem we want to find a reason for $S \models \varphi \wedge \neg\psi$ and a cause for $\neg\varphi \wedge \psi$ which are similar, while in the second problem we want to find a minimal modification to S that would guarantee that $\neg\varphi \wedge \psi$ holds.

We investigate our problems more closely in the setting of propositional logic (PL), assuming for simplicity that the output formulas are in conjunctive normal form. We first show that our definitions indeed find contrasts between the input formulas. For example, for the global contrastive explanation problem, we show that each clause C of the output formula θ is a *weak contrast* between φ and ψ , meaning φ entails C while ψ does not. We also show a link between our definitions and previous work on contrastive explanations: if the set S in the input of the counterfactual difference problem defines an assignment, then the output corresponds to a cardinality-minimal CXp [9]. As a by-product, we show that both of our counterfactual problems output partial assignments up to equivalence when presented with partial assignment inputs.

We also study the computational complexity of our problems in the case of PL. We first observe that the contrastive explanation problem, the global contrastive problem and the minimal separator problem are all Σ_2^P -complete. We then show that certain natural variants of our counterfactual problems are also Σ_2^P -complete. Finally, we provide a prototypical implementation of our problems using Answer Set Programming, which given the complexity of the underlying problems, is an adequate computation formalism [6]. We use this implementation to demonstrate how our problems work in practice via three case studies.

Related Work We now present related work on contrastive explanation within the broader context of logic-based explainable AI. It was originally for *local explanation*, namely explaining why a classifier makes a certain decision for a *given* input instance. As summarized in several works [9,13,4], local explanation answers one of the following two questions: Q1. (Why): what minimal aspects of an instance guarantee the actual decision? Q2. (Why not): what minimal changes to an instance result in a different decision? An answer to Q1 is nowadays commonly called a *sufficient reason* [5] or an AXp (short for abductive explanation) [10]. A sufficient reason is a minimal subset of its features' values s.t. changing any other feature values will not change the classification. In PL, a sufficient reason coincides with a *prime implicant* of the Boolean classifier φ which is *locally true* in the given instance. Therefore the concept was first introduced in the literature under the name PI-explanation [16]. For Barcelo et al. [2], a sufficient reason itself need not be minimal, and it is called a *minimal/minimum sufficient reason*, if it satisfies the subset-/cardinality-minimality requirement respectively. We refer to these notions as instances of *direct explanation*, for lack of a better term.

Dually, an answer to Q2 is commonly referred to as either a *contrastive explanation* (CXp) [9], a *counterfactual explanation* [17] or a necessary reason [4]. In all cases it is a subset-minimal part of the instance, changing the values of which results in a different classification. The duality of AXp and CXp is well-established from a logical viewpoint via minimal hitting sets [9]. Namely, an AXp is a subset-minimal intersection of every CXp of the instance and vice versa.

Similarly to before, a shift from subset-minimality to cardinality-minimality brings us to the notion of *minimum change required* by Barcelo et al. [2].

We now move from local to global explanation. Note that an explanation can be called ‘global’ in either *conditional* or *categorical* sense, namely either 1) it is a local explanation *if an input instance satisfies it*, or 2) it explains the whole classifier. In the former sense, the duality between AXp and CXp extends naturally to *global AXp* and *counterexamples* [13]. In propositional logic, they coincide with prime implicants and negated prime implicants. An example of a global direct explanation in the latter sense is the (shortest) prime DNF expression of Boolean classifiers [11]. However, here one cannot obtain a corresponding contrastive explanation in terms of the duality, because the contrastive explanations referred to so far are by nature localized to the actual case, viz., they are counterfactual. One must define a categorically global contrastive explanation in a genuinely global manner, which is a key contribution of our paper.

Besides the aforementioned ones, recently Bassan et al. [3] define *global sufficient reason* as the set of features which is a sufficient reason for all instances. Similarly, *global contrastive reason* is “a subset of features that may cause a misclassification for any possible input”. This notion of a global contrastive explanation differs significantly from ours, which aims to describe essentially all the differences between two classifiers.

2 Preliminaries

We will formulate our various explanation problems for an arbitrary logic. For the purposes of this work, a **logic** \mathcal{L} is a tuple $L = (\mathcal{F}, \models, \text{sat})$, where \mathcal{F} is a set of formulas, $\models \subseteq \mathcal{P}(\mathcal{F}) \times \mathcal{F}$ is a binary logical consequence relation and $\text{sat} \subseteq \mathcal{F}$ is a unary satisfiability relation. Instead of $\varphi, \psi \in \mathcal{F}$ and $\{\varphi\} \models \psi$ we will write $\varphi, \psi \in \mathcal{L}$ and $\varphi \models \psi$ for simplicity. If for $S, S' \subseteq \mathcal{F}$ we have $S \models \psi$ for every $\psi \in S'$ and $S' \models \varphi$ for every $\varphi \in S$, then we say that S and S' are equivalent, denoted $S \equiv S'$. We formulate our definitions for logics \mathcal{L} with classical conjunction \wedge and classical negation \neg with the usual semantics. These connectives are not necessary but they help to keep our definitions clean and readable.

As the logics that we consider are two-valued, the above framework works best when modeling classifiers for binary classification. To model classifiers with more than two classes, we can assign to each class c a formula φ_c which is satisfied precisely by the inputs classified as c . We call such a formula a **class formula**. This approach has been used, for example, in [4].

In addition to our very general definitions, we will also consider the important special case of propositional logic. Let τ be a set of **proposition symbols** called a **vocabulary**. The set $\text{PL}[\tau]$ of formulas of **propositional logic** over τ is generated by the grammar $\varphi ::= \perp \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi$, where $p \in \tau$. We define $\top := \neg\perp$ as an abbreviation.

A function $s : \sigma \rightarrow \{0, 1\}$, where $\sigma \subseteq \tau$, is called a **partial τ -assignment**. When $\sigma = \tau$, we say that s is a **τ -assignment**. We often identify (partial) assignments with conjunctions of literals in the obvious way. The semantics of

$\text{PL}[\tau]$ is defined as usual, i.e., for a τ -assignment s we define $s \not\models \perp$ always, $s \models p$ if $s(p) = 1$, $s \models \neg\psi$ if $s \not\models \psi$, $s \models \psi \wedge \theta$ if $s \models \psi$ and $s \models \theta$, and finally $s \models \psi \vee \theta$ if $s \models \psi$ or $s \models \theta$. Given $S \subseteq \text{PL}[\tau]$ and a (partial) τ -assignment s , we say that S **defines** s , if $S \equiv s$. If $S = \{\varphi\}$, we write φ defines s rather than $\{\varphi\}$ defines s .

Assignments that satisfy a formula are also called its **models**. For two formulas $\varphi, \psi \in \text{PL}[\tau]$, we use the notation $\varphi \models \psi$ for **logical consequence**, that is $\varphi \models \psi$ if for all assignments s , $s \models \varphi$ implies $s \models \psi$. We extend this notation also to sets $S, T \subseteq \text{PL}[\tau]$ of formulas, defining $S \models T$ if $\bigwedge S \models \bigwedge T$.

Formulas of the form p or $\neg p$, where p is a proposition symbol, are called **literals**. The **dual** $\bar{\ell}$ of a literal ℓ is defined as $\bar{p} = \neg p$ and $\overline{\neg p} = p$. A disjunction of literals is called a **clause**. A formula $\varphi \in \text{PL}[\tau]$ is in **conjunctive normal form**, or CNF, if φ is a conjunction of clauses. We will often denote clauses as sets of literals and CNF-formulas as sets of clauses. For example, if a CNF-formula φ has the clause $\neg p \vee q$, we might write $\{\neg p, q\} \in \varphi$. We denote the set of all CNF formulas $\varphi \in \text{PL}[\tau]$ by $\text{CNF}[\tau]$.

Formula size is a key notion in our work. We do not define formula size in the general case as many different definitions could make sense depending on the logic considered. We will, however, consider the size of propositional formulas in practice so we define it here. We define the **size** of a formula $\varphi \in \text{PL}[\tau]$ simply to be number of occurrences of proposition symbols. More formally, we define it recursively as follows: $\text{size}(\perp) = 0$ and $\text{size}(p) = 1$, for a proposition symbol p ; $\text{size}(\neg\psi) = \text{size}(\psi)$; $\text{size}(\psi_1 \wedge \psi_2) = \text{size}(\psi_1 \vee \psi_2) = \text{size}(\psi_1) + \text{size}(\psi_2)$.

3 Generalized Contrastive Explanation

In this section we introduce several natural problems that concern contrastive explanations. We start by defining problems focused on explaining differences either between instances (objects) or between properties. After this we consider problems that deal with the problem of explaining why a given instance has a particular property. In our logic-based framework instances are represented as sets of formulas while properties correspond to single formulas.

3.1 Comparison Explanations

We begin with a problem, where we have two different instances S and S' at hand, and we want to know why one of them satisfies φ while the other satisfies ψ . In practice, the two instances could seem very similar to us, leading to the question of key differences that explain their different properties φ and ψ . Van Bouwel and Weber [19] call this question an O-contrast, also cited by Miller [14].

Definition 1. *The **contrastive explanation problem** is defined as follows.*

Input: A tuple (S, S', φ, ψ) , where $S, S' \subseteq \mathcal{L}$ are finite sets and $\varphi, \psi \in \mathcal{L}$.

Output: A triple (θ, θ', χ) , where $\theta, \theta', \chi \in \mathcal{L}$ have the following properties.

1. $S \models \theta \wedge \chi \models \varphi \wedge \neg\psi$ and $S' \models \theta' \wedge \chi \models \neg\varphi \wedge \psi$,
2. $\text{size}(\theta) + \text{size}(\theta') + \text{size}(\chi)$ is minimal,

3. as a secondary optimization criterion, $\text{size}(\chi)$ is maximal.

If no such triple exists, output **error**.

The output contains the contrast formulas θ and θ' as well as a shared context formula χ . The essential part of the output are the differentiating formulas θ and θ' . The three conditions can be motivated as follows. Condition 1 ensures that $\theta \wedge \chi$ and $\theta' \wedge \chi$ serve as explanations for why S and S' satisfy $\varphi \wedge \neg\psi$ and $\neg\varphi \wedge \psi$, respectively. Condition 2 requires these explanations to be minimal to avoid including irrelevant information. Condition 3 enforces similarity between the explanations by maximizing the common context formula χ , as it is desirable to shift content from θ and θ' into χ .

In the above definition we are asking “Why $S \models \varphi \wedge \neg\psi$ and $S' \models \neg\varphi \wedge \psi$?” as opposed to “Why $S \models \varphi$ and $S' \models \psi$?”. The reason we do this is that, in accordance to Lipton [12], we allow the formulas φ and ψ to be compatible, but we seek explanations which entail that only one of them is true. We will follow this approach throughout the paper.

Example 1. We give a concrete example of Definition 1 in the case where \mathcal{L} is propositional logic. Consider the propositional vocabulary $\tau = \{p, q, r\}$. Let φ be a formula of $\text{PL}[\tau]$ which is satisfied by precisely those assignments which map exactly two propositional symbols to 1. Furthermore, let ψ be a formula of $\text{PL}[\tau]$ which is satisfied by precisely those assignments which map exactly one propositional symbol to 1. Note that $\varphi \models \neg\psi$ and $\psi \models \neg\varphi$, whence $\varphi \wedge \neg\psi \equiv \varphi$ and $\neg\varphi \wedge \psi \equiv \psi$. Now, let $S := \{p, q, \neg r\}$ and $S' := \{p, \neg q, \neg r\}$. The following triple is a possible solution: $\theta := q$, $\theta' := \neg q$ and $\chi := p \wedge \neg r$. ■

Consider the special case of the contrastive explanation problem, where $S = \{\varphi \wedge \neg\psi\}$ and $S' = \{\neg\varphi \wedge \psi\}$. Now we are asking for a contrast between all of the models of φ and ψ , with no limitation to a more specific locality S or S' . Thus we are asking “What is the difference between φ and ψ ?” We call this case the global contrastive explanation problem and define it next separately.

Definition 2. The *global contrastive explanation problem* is defined as follows.

Input: A pair (φ, ψ) , where $\varphi, \psi \in \mathcal{L}$.

Output: A triple (θ, θ', χ) , where $\theta, \theta', \chi \in \mathcal{L}$ have the following properties.

1. $\theta \wedge \chi \equiv \varphi \wedge \neg\psi$ and $\theta' \wedge \chi \equiv \neg\varphi \wedge \psi$,
2. $\text{size}(\theta) + \text{size}(\theta') + \text{size}(\chi)$ is minimal,
3. as a secondary optimization criterion, $\text{size}(\chi)$ is maximal.

Example 2. Consider again the formulas φ and ψ from Example 1. Restricting our output formulas to CNF-formulas for readability, the following triple is a solution to the global contrastive explanation problem.

$$\begin{aligned}\theta &:= (p \vee r) \wedge (q \vee r) \wedge (p \vee q) \\ \theta' &:= (\neg p \vee \neg r) \wedge (\neg q \vee \neg r) \wedge (\neg p \vee \neg q) \\ \chi &:= (p \vee q \vee r) \wedge (\neg p \vee \neg q \vee \neg r)\end{aligned}$$

These formulas tell us the following. First, θ says that in every model of $\varphi \wedge \neg\psi \equiv \varphi$ at least two propositional symbols are true. Secondly, θ' says that in every model of $\neg\varphi \wedge \psi \equiv \psi$ at least two propositional symbols are false. Finally, χ tells us that in models of $\varphi \vee \psi$ one propositional symbol is true and one is false. ■

Another notion related to contrastivity is that of a separator. A separator of φ from ψ is a property that all models of φ and no models of ψ have. A separator can thus also be seen as a different answer to the question “What is the difference between φ and ψ ?” The next problem asks for a minimal separator of φ from ψ .

Definition 3. The *minimal separator problem* is defined as follows.

Input: A pair (φ, ψ) , where $\varphi, \psi \in \mathcal{L}$.

Output: A formula $\theta \in \mathcal{L}$ such that $\varphi \models \theta$, $\psi \models \neg\theta$ and $\text{size}(\theta)$ is minimal. If no such θ exists, output **error**.

The global contrastive explanation problem can be seen as giving “all” separators between φ and ψ , or at least enough to achieve equivalent formulas, while the minimal separator problem only gives one minimal separator. In terms of the natural language question “What is the difference between φ and ψ ?”, giving all differences or a single difference could both be considered reasonable answers.

Example 3. Consider again the formulas φ, ψ in Example 1. Clearly $\varphi \models \neg\psi$. The formula $(p \vee r) \wedge (q \vee r) \wedge (p \vee q)$ is a minimal separator between φ and ψ . ■

3.2 Counterfactual Explanations

We have seen above how the contrastive explanation problem answers the question “Why does S satisfy φ while S' satisfies ψ ?” We now turn our attention to the case, where we have only one instance S at hand and we are asking about the classification of that instance. The question here is “Why does S satisfy φ and not ψ ?”. Van Bouwel and Weber [19] call this question a P-contrast, also cited by Miller [14]. We define two problems that modify the contrastive explanation problem in different ways to answer this question.

Our first problem is a straightforward modification of Definition 1, where we simply leave S' to be existentially quantified. The condition $S' \models \theta' \wedge \chi \models \neg\varphi \wedge \psi$ is then reduced to the form $\theta' \wedge \chi \models \neg\varphi \wedge \psi$ since S' itself is not actually needed for the output. The definition is as follows.

Definition 4. The *counterfactual contrastive explanation problem* is defined as follows.

Input: A tuple (S, φ, ψ) , where $S \subseteq \mathcal{L}$ is a finite set and $\varphi, \psi \in \mathcal{L}$.

Output: A triple (θ, θ', χ) , where $\theta, \theta', \chi \in \mathcal{L}$ have the following properties.

1. $S \models \theta \wedge \chi \models \varphi \wedge \neg\psi$ and $\theta' \wedge \chi \models \neg\varphi \wedge \psi$,
2. $\theta' \wedge \chi$ is satisfiable iff S is satisfiable,
3. $\text{size}(\theta) + \text{size}(\theta') + \text{size}(\chi)$ is minimal,
4. as a secondary optimization criterion, $\text{size}(\chi)$ is maximal.

If no such triple exists, output **error**.

The above definition places emphasis on minimal reasons for why instances satisfy properties. Another way to think about contrastive explanations is to consider minimal changes required to the input instance in order to achieve the desired outcome. This can be seen in the literature in the concept of CXp [9]. It turns out that minimal reasons for satisfaction and minimal changes to achieve satisfaction do not always coincide. This difference between reasons, or why-questions, and changes, or what-is-the-difference-questions motivates our next definition. The intuitive question here is “What is the difference between S that satisfies φ and the closest S' that instead satisfies ψ ?”

Definition 5. The *counterfactual difference problem* is defined as follows.

Input: A tuple (S, φ, ψ) , where $S \subseteq \mathcal{L}$ is a finite set and $\varphi, \psi \in \mathcal{L}$.

Output: A triple (θ, θ', χ) , where $\theta, \theta', \chi \in \mathcal{L}$ have the following properties.

1. $S \equiv \theta \wedge \chi \models \varphi \wedge \neg\psi$ and $\theta' \wedge \chi \models \neg\varphi \wedge \psi$,
2. $\theta' \wedge \chi$ is satisfiable iff S is satisfiable,
3. $\text{size}(\theta) + \text{size}(\theta') + \text{size}(\chi)$ is minimal,
4. as a secondary optimization criterion, $\text{size}(\chi)$ is maximal.

If no such triple exists, output **error**.

Remark 1. Note that for both Definitions 4 and 5, if $\psi \models \varphi$ and S is satisfiable, then the output is **error**. Here the models of ψ are included in those of φ and thus $\neg\varphi \wedge \psi$ is a contradiction. The input can be repaired by replacing φ with $\varphi \wedge \neg\psi$, since $\neg(\varphi \wedge \neg\psi) \wedge \psi \equiv \psi$. This makes the two input formulas separate and preserves the original question: “Why does S satisfy φ but not ψ ?”

Example 4. Assume we have two propositional classifiers, φ and ψ , trained on some data about seabirds. The classifier $\varphi = \text{beak_pouch}$ classifies a bird as a pelican if the bird has a distinctive pouch on the underside of its beak. The classifier $\psi = \neg\text{beak_pouch} \wedge \text{small} \wedge ((\text{white_body} \wedge \text{webbed_feet}) \vee \text{grey_wing})$ classifies a bird as a seagull if it has no beak pouch, is less than 1 meter in size and either has white plumage on its body and webbed feet, or a grey wing. Note that these are not complete descriptions of these types of birds but rather classification criteria that could have been extracted from a dataset.

Let $S = \{\text{beak_pouch}, \neg\text{small}, \text{white_body}, \text{webbed_feet}, \neg\text{grey_wing}\}$ be a description of a bird. We want to know why this bird was classified as a pelican and not as a seagull. Starting with Definition 5, the output in this case is $\theta = \text{beak_pouch}$, $\theta' = \text{small}$, $\chi = (\neg\text{beak_pouch} \vee \neg\text{small}) \wedge \text{white_body} \wedge \text{webbed_feet} \wedge \neg\text{grey_wing}$. We can read the solution as “This bird has a beak pouch and is large, so it’s a pelican. If it had no beak pouch and was small, it would instead be a seagull.” Note how all attributes of the bird are listed in the formula $\theta \wedge \chi$ even though beak_pouch suffices to classify the bird as a pelican.

The output of Definition 4 is $\theta = \text{beak_pouch}$, $\theta' = \neg\text{beak_pouch} \wedge \text{small} \wedge \text{grey_wing}$, $\chi = \top$. This would be read as “This bird has a beak pouch so it’s a pelican. If it had no beak pouch, was small and had a grey wing, it would instead

be a seagull.” This time only the beak pouch is listed in $\theta \wedge \chi$ as it suffices to classify the bird as a pelican.

For another difference between the definitions, note that out of the two options provided by the classifier ψ , Definition 4 has chosen the one with the grey wing. Another option in the search space would have been $\theta = \text{beak_pouch}$, $\theta' = \neg \text{beak_pouch} \wedge \text{small}$, $\chi = \text{white_body} \wedge \text{webbed_feet}$. Out of these two, the grey wing option was chosen because the reasons $\theta \wedge \chi$ and $\theta' \wedge \chi$ given for why the bird was a pelican or a seagull were shorter in the first option. This illustrates the fact that Definition 4 is not concerned with minimal differences between the input and the counterfactual case, but rather minimal differences between minimal reasons for the classifications of the input and the counterfactual case. ■

4 The Case of Propositional Logic

In this section, we investigate our problems more closely in the setting of propositional logic. We show that our problems indeed find contrasts between the inputs in a semantic sense. We also establish a formal link between our definitions and existing work on contrastive explanations. We conclude with a study of the computational complexity of our problems.

4.1 Contrasts and Likenesses

We start by defining some notions that compare two formulas φ and ψ .

Definition 6. Let φ, ψ, θ be formulas of a logic \mathcal{L} .

1. θ is a **weak** (φ, ψ) -**contrast** if $\varphi \wedge \neg\psi \models \theta$ and $\neg\varphi \wedge \psi \not\models \theta$.
2. θ is a **strong** (φ, ψ) -**contrast** if $\varphi \wedge \neg\psi \models \theta$ and $\neg\varphi \wedge \psi \models \neg\theta$.
3. θ is a (φ, ψ) -**likeness** if $\varphi \wedge \neg\psi \models \theta$ and $\neg\varphi \wedge \psi \models \theta$.

Strong contrasts are properties that all models of φ (or more precisely, $\varphi \wedge \neg\psi$), and none of the models of ψ , have. Weak contrasts, on the other hand, are properties that all models of, say, φ have, but not all models of ψ do. Likenesses are properties that models of both formulas have.

For global contrastive explanations, the output formulas correspond to contrasts and likenesses in a nice way. Let (θ, θ', χ) be an output of the global contrastive explanation problem (Definition 2). We have $\theta \wedge \chi \wedge \theta' \equiv \varphi \wedge \neg\psi \wedge \neg\varphi \wedge \psi \rightarrow \perp$, which means $\theta \wedge \neg\varphi \wedge \psi \rightarrow \perp$, namely $\neg\varphi \wedge \psi \rightarrow \neg\theta$, i.e. θ is always a strong (φ, ψ) -contrast. By symmetry θ' is a strong (ψ, φ) -contrast. It is also easy to see that χ is a (φ, ψ) -likeness. The following result further shows that if we assume the output formulas are in CNF, then the individual clauses of θ and θ' are all weak contrasts whereas the clauses of χ are likenesses of φ and ψ .

Theorem 1. Let $\varphi, \psi \in \text{PL}[\tau]$ and let (θ, θ', χ) be the output of the global contrastive explanation problem with input (φ, ψ) . Further assume that θ , θ' and χ are in CNF. Then (1) each clause of θ is a weak (φ, ψ) -contrast, (2) each clause of θ' is a weak (ψ, φ) -contrast, and (3) each clause of χ is a (φ, ψ) -likeness.

Proof. Condition 2 of the problem means that it is generally more efficient in terms of formula size to list likenesses in χ rather than θ and θ' . See the full version for the proof. \square

Remark 2. Minimizing $\text{size}(\theta) + \text{size}(\theta') + \text{size}(\chi)$ is not the only conceivable minimality condition for the formulas $\theta \wedge \chi$ and $\theta' \wedge \chi$. If one thinks of these as two formulas to be minimized, then another natural condition could be $\text{size}(\theta \wedge \chi) + \text{size}(\theta' \wedge \chi)$. For the global contrastive explanation problem, changing to this alternate condition would forfeit the property of Theorem 1, but gain a property, where clauses that are strong contrasts can be easily identified from the output.

We also note that the same kind of property holds for the contrastive explanation problem. The difference is that here, the contrasts are found between S and S' rather than φ and ψ .

Theorem 2. *Let $\varphi, \psi \in \text{PL}[\tau]$ and let (θ, θ', χ) be the output of the contrastive explanation problem with input (S, S', φ, ψ) . Further assume that θ, θ' and χ are in CNF. Then (1) each clause of θ is a weak (S, S') -contrast, (2) each clause of θ' is a weak (S', S) -contrast, and (3) each clause of χ is a (S, S') -likeness.*

Proof. We first note that since $S \models \varphi \wedge \neg\psi$ and $S' \models \neg\varphi \wedge \psi$, we have $S \wedge \neg S' \equiv S$. Now for a clause $C \in \theta$, we have $S \models \theta \models C$ and we can use the same arguments as in the proof of Theorem 1 to prove the claim. \square

Note that Theorem 2 also works for the counterfactual problems if we consider an existentially quantified S' to be implicitly present in the definitions.

4.2 Link to Existing Contrastive Explanations

In this subsection we link the counterfactual difference problem to existing notions in the literature, such as CXp. For example, we show that if the input S defines an assignment and output formulas θ, θ' and χ are given in CNF, then the output gives the minimal changes required to flip the truth values of φ and ψ . The difference from CXp is that our definition gives a cardinality-minimal solution rather than a subset-minimal one.

We start with the following theorem which shows that in the special case of the counterfactual difference problem where S defines a partial assignment, the optimal solutions also define partial assignments. Example 4 shows that in general we cannot require χ to even define a partial assignment.

Theorem 3. *Let $\varphi, \psi \in \text{PL}[\tau]$ and let $S \subseteq \text{PL}[\tau]$ define a partial assignment. Let (θ, θ', χ) be the output of the counterfactual difference problem (S, φ, ψ) . Further assume that θ, θ' and χ are in CNF. Then (1) the formulas θ and θ' are partial assignments and (2) the formulas $\theta \wedge \chi$ and $\theta' \wedge \chi$ define partial assignments. The same also holds for the counterfactual contrastive explanation problem.*

Proof. The proofs for both problems are similar. We assume that an output (θ, θ', χ) does not satisfy the claims and then prune or move clauses to find an alternative output $(\theta_*, \theta'_*, \chi_*)$ with smaller total size that does satisfy the claims. This contradicts condition 3 of the problems. See the full version for the proof. \square

In the case where S defines a τ -assignment, we get a stronger guarantee on optimal outputs for the counterfactual difference problem.

Theorem 4. *Let $\varphi, \psi \in \text{PL}[\tau]$, let $S \subseteq \text{PL}[\tau]$ define a τ -assignment and let (θ, θ', χ) be the output of the counterfactual difference problem with input (S, φ, ψ) . Further assume that θ, θ' and χ are in CNF. Now the formulas $\theta \wedge \chi$ and $\theta' \wedge \chi$ define τ -assignments.*

Proof. We know that $\theta \wedge \chi$ defines an assignment and $\theta' \wedge \chi$ defines a partial assignment. We show how a proposition missing from this partial assignment could be included with formulas that are more optimal in terms of conditions 3 and 4 of the problem. See the full version for the proof. \square

The above theorem does not hold for the counterfactual contrastive explanation problem. This is to be expected as the condition $S \models \theta \wedge \chi$ is not sufficient to force $\theta \wedge \chi$ to define an assignment, again highlighting the fact that the counterfactual contrastive explanation problem is concerned with comparing *reasons* for the formulas φ and ψ rather than assignments that satisfy them.

To state the last result of this section, we define the notation $s \triangle \lambda = (s \setminus \lambda) \cup \{\bar{\ell} \mid \ell \in \lambda\}$, where s and λ are viewed as sets of literals with $\lambda \subseteq s$. This result is of particular interest, as when the second input formula ψ is the negation $\neg\varphi$ of the first, the set λ of propositions to be flipped is a cardinality minimal CXp [9].

Theorem 5. *Let s be a τ -assignment and let $\varphi, \psi \in \text{PL}[\tau]$ such that $\neg\varphi \wedge \psi$ is satisfiable. Assume that $s \models \varphi \wedge \neg\psi$ and let (θ, θ', χ) be the output of the counterfactual difference problem for PL with input (S_s, φ, ψ) . Further assume that θ, θ' and χ are in CNF. Now $\theta' \wedge \chi$ defines a τ -assignment s' such that $s' = s \triangle \lambda$ for a cardinality-minimal set λ of literals with $s \models \lambda$, $s \triangle \lambda \models \neg\varphi \wedge \psi$.*

Proof. We know from Theorem 4 that $\theta' \wedge \chi$ defines a τ -assignment s' . We show that given two candidates for s' , the one with less differences compared to s is the preferred candidate due to the formula size condition 3 of the problem. See the full version for the proof. \square

4.3 Computational Complexity

We proceed to study the computational complexity of our problems for propositional logic. As per usual, we study the complexity of the related decision problems: instead of finding a minimal solution to the given problem, we want to decide whether there is a solution of size at most k , where k is part of the input.

We start with global contrastive explainability problem. Recall that in the propositional setting we require the output formulas to be CNF-formulas.

Theorem 6. *The global contrastive explanation problem is Σ_2^P -complete.*

Proof. We will give a simple reduction from the minimal formula size problem for CNF-formulas, which was proved to be Σ_2^P -complete in [18]. Let (φ, k) be an input to the latter problem. Consider the following input (\perp, φ, k) to the global contrastive explainability problem. If (θ, θ', χ) is a witness for this input, then $\theta \equiv \perp$ and $\theta' \wedge \chi \equiv \varphi$. It is now easy to see that the output of (φ, k) is **yes** iff the output of (\perp, φ, k) is **yes**. \square

For the next hardness result we use a result from [11] which states that the so-called **local explainability problem** for PL is Σ_2^P -complete. In this problem the input is a tuple (s, φ, k) , where s is an assignment, such that $s \models \varphi$ and the goal is to determine whether there exist a formula ψ of PL such that $\text{size}(\psi) \leq k$ and $s \models \psi \models \varphi$. The following result demonstrates that this problem is a special case of the contrastive explainability problem and the separability problem.

Theorem 7. *Both the contrastive explanation problem and the minimal separator problem for PL are Σ_2^P -complete.*

Proof. Upper bound is clear. For the lower bounds, consider an instance (s, φ, k) of the local explainability problem. Let φ_s be a conjunction of literals which defines s . For contrastive explainability problem the hardness follows from the observation that $(\varphi_s, \perp, \varphi, \perp, k)$ is an instance of the contrastive explainability problem for which the output is **yes** iff the output of $(s, \varphi, k, 1)$ is **yes**. In the right-to-left direction we use a result from [11] which shows that if there exist a formula ψ with $\text{size}(\psi) \leq k$ and $s \models \psi \models \varphi$, then there is also one which is a conjunction of literals. For separability problem we get the hardness by considering instances of the form $(\varphi_s, \neg\varphi, k)$. \square

We move on to consider counterfactual contrastive and difference problems. Theorem 3 shows that if the input S defines a partial assignment, then the formulas $\theta \wedge \chi$ and $\theta' \wedge \chi$ from the output (θ, θ', χ) also define partial assignments. Thus it is natural to define the **simplified counterfactual contrastive explanation problem** and the **simplified counterfactual difference problem** by modifying Definitions 4 and 5 as follows. First, we require that S defines a partial assignment. Secondly, the output formulas θ , θ' and χ are required to be partial assignments.

Our next result shows that already the simplified problems are Σ_2^P complete. See the full version for the proof. We leave the complexity of the non-simplified problems for future work.

Theorem 8. *The simplified counterfactual contrastive explanation problem and the simplified counterfactual difference problem are Σ_2^P -complete.*

5 Implementation and Case Studies

In this section we outline our Answer Set Programming (ASP) based implementation for selected explanation problems, followed by case studies where we use it to demonstrate how our problems work on real-world instances.

Implementation Due to the computational complexity of the considered problems, obtaining a polynomial time dedicated algorithm seems unlikely. However, the inherent complexity closely matches the one of ASP [6] and it is thus natural to implement our notions in this formalism to obtain a prototypical implementation.

We thus implemented Definitions 2, 4, and 5 in ASP. The encodings as well as a Python script for reproducing the case studies can be found online at https://github.com/tlyphed/general_contrastive_exp. Due to length constraints we cannot cover the ASP encoding in detail but we will give a brief overview.

The idea behind the encodings is to guess the output formulas θ, θ', χ as CNF and check that the required entailments and equivalences hold. Those checks are done using the *saturation technique* [6], which is an encoding method for expressing **coNP** problems in ASP. The criteria regarding the sizes of the formulas are then formulated using weak constraints, which essentially give preference to formulas which adhere to those criteria.

Case Studies We used our implementation to compute contrastive explanations for decision trees that were obtained from three real-world classification datasets. The datasets that we used were Iris [7], Wine [1] and Glass [8]. We used the `scikit-learn` Python library [15] for training the classifiers. For all datasets we used 80% of the data for training and 20% for testing. To keep the learning task simple, we fine-tuned only the depth of the decision trees, selecting the smallest depth that achieved the highest accuracy on the test data.

Having trained the decision trees, we proceeded as follows. For each dataset we picked two classes c, c' at random and then used the learned decision tree \mathcal{T} to form class formulas for c, c' . The propositional symbols in these class formulas correspond to pivots used by \mathcal{T} . We also picked a random instance from the test data that was classified as c by \mathcal{T} . We then ran our implementations for global contrastive explanation problem (GCE), counterfactual contrastive explanation problem (CCE) and counterfactual difference problem (CD) using the class formulas and the instance as inputs. For CCE and CD, we additionally constrained the output formulas to be partial assignments. This choice is motivated by Theorems 3 and 4, since in our case studies S (the instance) corresponds to a conjunction of literals. The results are summarized in Table 1.

We make some general remarks about Table 1. For GCE we see both simple and more complex likenesses. For example, in the case of Glass, the optimal output had no common clauses for the two class formulas. In each case we see that CCE and CD have chosen the same propositions to flip, although we know by Example 4 that this is not necessary. The CCE outputs are more informative here since they also leave out unnecessary propositions from the explanations.

We examine the formulas in more detail for the Iris dataset. Each instance is an iris flower and the goal is to classify them into one of three species: *setosa*, *versicolor*, or *virginica*. We formed class formulas for *versicolor* and *virginica*. From GCE we see e.g. that the class formulas have $\neg p_1$ in common, where

$$p_1 := \text{“petal length is } \leq 2.45\text{cm”}.$$

Table 1. Summary of our case studies. Note that we have Booleanized the instances using the pivots learned by the corresponding decision tree.

Case	GCE	CCE	CD
dataset: Iris	$\theta : (\neg p_2 \vee p_3) \wedge (p_2 \vee p_4)$	$\theta : p_3$	$\theta : p_3$
depth: 4	$\theta' : (\neg p_2 \vee \neg p_3) \wedge (p_2 \vee \neg p_4)$	$\theta' : \neg p_3$	$\theta' : \neg p_3$
instance: $\{\neg p_1, p_2, p_3, p_4\}$	$\chi : \neg p_1$	$\chi : \neg p_1 \wedge p_2$	$\chi : \neg p_1 \wedge p_2 \wedge p_4$
dataset: Wine	$\theta : (p_1 \vee \neg p_4) \vee (p_1 \vee p_5)$	$\theta : p_1$	$\theta : p_1 \wedge \neg p_4$
depth: 3	$\theta' : \neg p_1 \wedge p_4$	$\theta' : \neg p_1 \wedge p_4$	$\theta' : \neg p_1 \wedge p_4$
instance: $\{p_1, p_2, p_3, \neg p_4, p_5\}$	$\chi : (\neg p_1 \vee p_3) \wedge (\neg p_1 \vee p_2)$	$\chi : p_2 \wedge p_3$	$\chi : p_2 \wedge p_3 \wedge p_5$
dataset: Glass	$\theta : (\neg p_1 \vee p_2) \wedge (p_1 \vee \neg p_5)$		$\theta : p_2 \wedge \neg p_4$
depth: 3	$\wedge(\neg p_1 \vee \neg p_3) \wedge (p_1 \vee \neg p_7)$	$\theta : p_2$	$\theta' : \neg p_2 \wedge p_4$
instance: $\{p_1, p_2, \neg p_3, \neg p_4, \neg p_5, p_6, p_7\}$	$\theta' : (p_1 \vee p_5) \wedge (p_1 \vee \neg p_6)$ $\wedge(\neg p_1 \vee p_4) \wedge (\neg p_1 \vee \neg p_2)$ $\chi : \top$	$\theta' : \neg p_2 \wedge p_4$ $\chi : p_1 \wedge \neg p_3$	$\chi : p_1 \wedge \neg p_3 \wedge \neg p_5$ $\wedge p_6 \wedge p_7$

So the class formulas for versicolor and virginica both agree that petal length should be more than 2.45cm. From CCE and CD, we see that the instance $\{\neg p_1, p_2, p_3, p_4\}$ was classified as versicolor, because $p_3 \wedge \neg p_1 \wedge p_2$, where

$$p_2 := \text{“petal length is } \leq 4.75\text{cm”}$$


and

$$p_3 := \text{“petal width is } \leq 1.65\text{cm”}.$$

So the explanation is that the flower has petal length between 2.45cm and 4.75cm and petal width at most 1.65cm. If the petal length had been more than 4.75cm, the flower would have been classified as virginica.

6 Conclusion

In this work, we introduced a logic-based framework for contrastive explanations that formalizes various questions of the form “Why P but not Q ?”. Our framework encompasses both local and global contrastive explanations. We examined the theoretical properties of our definitions in detail in the important special case of propositional logic. Among other results, we showed that our framework subsumes a cardinality-minimal version of existing contrastive explanation approaches from [4,9]. In addition, we analyzed the computational complexity of the proposed problems and proved that, in the propositional setting, most of them are Σ_2^P -complete. Finally, we implemented our approach using Answer Set Programming and demonstrated that our problems produce useful explanations on real-world instances.

Acknowledgments. Antti Kuusisto and Miikka Vilander received funding from the Research Council of Finland projects *Explaining AI via Logic* (XAILOG) 345612 and *Theory of computational logics* 324435, 328987, 352419, 352420. Tobias Geibinger is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Logic and Computation at the TU Wien. This work has benefited from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101034440 (LogiCS@TUWien).  In addition, this research was funded in part by the Austrian Science Fund (FWF) [10.55776/COE12]. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.

The author names of this article are ordered based on alphabetical order.

References

1. Aeberhard, S., Forina, M.: Wine. UCI Machine Learning Repository (1992). <https://doi.org/10.24432/C5PC7J>
2. Barceló, P., Monet, M., Pérez, J., Subercaseaux, B.: Model interpretability through the lens of computational complexity. *Advances in neural information processing systems* **33**, 15487–15498 (2020)
3. Bassan, S., Amir, G., Katz, G.: Local vs. global interpretability: A computational complexity perspective. In: Forty-first International Conference on Machine Learning (2024)
4. Darwiche, A.: Logic for explainable AI. In: LICS. pp. 1–11 (2023). <https://doi.org/10.1109/LICS56636.2023.10175757>, <https://doi.org/10.1109/LICS56636.2023.10175757>
5. Darwiche, A., Hirth, A.: On the reasons behind decisions. In: 24th European Conference on Artificial Intelligence (ECAI 2020). *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 712–720. IOS Press (2020)
6. Eiter, T., Ianni, G., Krennwallner, T.: Answer set programming: A primer. In: Reasoning Web. Semantic Technologies for Information Systems. LNCS, vol. 5689, pp. 40–110. Springer (2009). https://doi.org/10.1007/978-3-642-03754-2_2
7. Fisher, R.A.: Iris. UCI Machine Learning Repository (1936). <https://doi.org/10.24432/C56C76>
8. German, B.: Glass Identification. UCI Machine Learning Repository (1987). <https://doi.org/10.24432/C5WW2P>
9. Ignatiev, A., Narodytska, N., Asher, N., Marques-Silva, J.: From contrastive to abductive explanations and back again. In: Baldoni, M., Bandini, S. (eds.) *AIXIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence*, Virtual Event, November 25–27, 2020, Revised Selected Papers. *Lecture Notes in Computer Science*, vol. 12414, pp. 335–355. Springer (2020). https://doi.org/10.1007/978-3-030-77091-4_21, https://doi.org/10.1007/978-3-030-77091-4_21
10. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: *Proceedings of the Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*. vol. 33, pp. 1511–1519 (2019)
11. Jaakkola, R., Janhunen, T., Kuusisto, A., Rankooh, M.F., Vilander, M.: Explainability via short formulas: the case of propositional logic with implementation. In: *Joint Proceedings of (HYDRA 2022) and the RCRA Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion*. *CEUR Workshop Proceedings*, vol. 3281, pp. 64–77 (2022)

12. Lipton, P.: Contrastive explanation. *Royal Institute of Philosophy Supplement* **27**, 247–266 (1990). <https://doi.org/10.1017/S1358246100005130>
13. Marques-Silva, J.: Logic-based explainability in machine learning. In: Bertossi, L.E., Xiao, G. (eds.) *Reasoning Web. Causality, Explanations and Declarative Knowledge - 18th International Summer School 2022*, Berlin, Germany, September 27–30, 2022, Tutorial Lectures. *Lecture Notes in Computer Science*, vol. 13759, pp. 24–104. Springer (2022). https://doi.org/10.1007/978-3-031-31414-8_2, https://doi.org/10.1007/978-3-031-31414-8_2
14. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267**, 1–38 (2019). <https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007>
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
16. Shih, A., Choi, A., Darwiche, A.: Formal verification of bayesian network classifiers. In: *International Conference on Probabilistic Graphical Models*. pp. 427–438. PMLR (2018)
17. Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *Ieee Access* **9**, 11974–12001 (2021)
18. Umans, C.: The minimum equivalent dnf problem and shortest implicants. *Journal of Computer and System Sciences* **63**(4), 597–611 (2001). <https://doi.org/https://doi.org/10.1006/jcss.2001.1775>
19. Van Bouwel, J., Weber, E.: Remote causes, bad explanations? *Journal for the Theory of Social Behaviour* **32**(4), 437 – 449 (2002). <https://doi.org/10.1111/1468-5914.00197>