European Journal of Operational Research 329 (2026) 24–41

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Continuous Optimization

Counterfactual explanations for linear optimization

Jannis Kurtz *, Ş. İlker Birbil, Dick den Hertog

University of Amsterdam, Amsterdam Business School, Plantage Muidergracht 12, Amsterdam, 1018 TV, Netherlands

ARTICLE INFO

Keywords:
Linear programming
Counterfactual explanations
Bilinear optimization
Inverse optimization

ABSTRACT

In recent years, the concept of counterfactual explanations (CE) has become increasingly important in understanding the inner workings of complex AI systems. In this paper, we introduce the idea of CEs in the context of linear optimization and propose, explain, and analyze three different classes of CEs: relative, weak, and strong. We discuss in which situation each type of CE is needed and examine the structure of the optimization problems that arise from considering them. By detecting and leveraging the underlying convex structure of the relative CE problem, we demonstrate that computing the relative CEs takes the same order of time as solving the original problems. We also address the computational challenges associated with weak and strong CE problems. To illustrate our findings, we present a case study with data sourced from the World Food Programme in which we calculate each type of CE. Finally, we conduct comprehensive numerical experiments using the NETLIB library to demonstrate that relative CE problems can be solved as quickly as solving the original linear optimization problem.

1. Introduction

As artificial intelligence (AI) continues to influence our daily lives, the need for interpretability and transparency increases. This need for comprehensive explanations has been accelerated partly by the legislative initiatives such as the General Data Protection Regulation, the European Union AI Act, and the US Blueprint for an AI Bill of Rights (EUR-Lex, 2016, 2021; OSTP, 2022). These regulations emphasize the necessity of providing clear and understandable explanations for automated systems, echoing society's demand for trustworthy AI and aligning with the *right for explanation* principle.

These developments have attracted the attention of the researchers in machine learning who have started to develop algorithms that pave the way for explainable AI (XAI) (Biran & Cotton, 2017). Among these efforts, the concept of counterfactual explanations (CEs) has emerged as one of the key approaches in XAI to understanding the inner workings of complex AI models (Maragno et al., 2022; Wachter et al., 2018). CEs aim to identify the (smallest) change in data that would lead to a desired model outcome. A canonical CE example is credit scoring, where a model predicts loan eligibility. If the model denies a loan for an individual, then it should also offer an explanation. For instance a CE might state: "If your annual salary was 1500 EUR higher and your account balance was 900 EUR higher, you would have been granted a loan".

While much attention is dedicated to the explanations of AI systems, only a few works tackle explainability of decisions stemming from the solutions of optimization problems (Aigner et al., 2024; Goerigk & Hartisch, 2023; Korikov & Beck, 2021, 2023; Korikov et al., 2021). These solutions play a pivotal role in diverse domains, ranging from logistics and finance to healthcare and engineering. While the mechanisms of the underlying optimization algorithms (*e.g.*, the simplex method) may be clear to optimization experts, this is not the case for individuals who are affected by the decisions of the algorithm; *e.g.*, doctors, nurses or patients in the health-care sector. While it is generally accepted in the AI community that the reasoning behind the evaluation of a neural network (which is basically performing matrix multiplications) can be seen as a black-box, evaluating the outcome of a huge number of simplex iterations or an exponentially sized branch-and-bound tree is not framed in the same way.

The significance of explanations in optimization becomes apparent as they offer advantages at various levels of application. First, they can be used to support individuals attempting to understand the reasoning behind optimization-driven decisions. Second, stakeholders, such as businesses and public authorities, are impacted by optimization results and can use explanations to get clear justifications for decisions that may have broad implications. Finally, the operations research analyst, who is responsible for setting up complex models, can greatly benefit from substantial insights into the complex interactions of variables and parameters. To this end, we propose to apply and extend the concept of CEs to linear optimization. As defined in Korikov et al. (2021), we obtain a CE by identifying the (smallest) change needed

* Corresponding author.
E-mail address: j.kurtz@uva.nl (J. Kurtz).

https://doi.org/10.1016/j.ejor.2025.06.016
Received 18 December 2024; Accepted 12 June 2025
Available online 18 July 2025

in the optimization model's parameters such that an optimal solution of the new problem fulfills a set of desired properties. One of the main drawbacks of the latter concept is the computational intractability of deriving such CEs in the general case where objective and constraint parameters may be changed, even for classical linear optimization problems. Hence, in this work we develop a new notion of CEs which is computationally tractable and can be solved in the same order of time as the underlying linear problem.

1.1. Examples

In this section, we describe in detail several motivating examples where counterfactual explanations are crucial.

Diet problem. In Peters et al. (2021, 2022) a linear optimization model is developed to optimize the food supply chain for the United Nations' World Food Programme (WFP). This model has been and is used for each project of the WFP, and has enabled WFP to feed millions of people. An important part of the model is the diet problem: Which food commodities should be included in the food basket such that all nutritional requirements are satisfied, while the costs are minimized? The food commodities can be purchased from different suppliers, either from local or international markets. Suppose that the optimal solution of the linear optimization model is such that only a limited amount of products of the optimal food basket is purchased from the local market. Since purchases from local markets are preferred (due to smaller transportation costs), the decision maker wants to know for which changes in the purchasing costs or nutrition values, the amount purchased from the local market increases by 5%. The corresponding counterfactual explanation provides insights on this question. In another situation it could happen that a supplier who is selling parts of the food commodities wants to sell more of a certain type of food (*e.g.*, Wheat) due to a large stock level. The supplier could ask the counterfactual question: "What are the minimal changes in my prices such that at least a certain amount of Wheat is purchased from me?"

A similar situation occurs in the mobile application Feed Calculator (Meijer & van Veluw, 2024), which is now used by thousands of small farmers in Africa and Asia to optimize the ingredients for the cattle feed. Each possible ingredient can be purchased at a certain local supplier. The core of this application is the diet model, which is a small linear optimization problem. The local suppliers could use counterfactual explanation for this linear optimization model to detect minimal changes to the costs or nutritional contents such that its food commodity becomes attractive for being purchased by local farmers.

Facility location problem. One of the Sustainable Development Goals of the United Nations is good geospatial accessibility to healthcare centers in low- and middle-income countries. Facility location approaches have been developed to optimize geospatial accessibility given a certain budget to build new centers (Krishnakumari et al., 2024). For local governments it is crucial to know, for example, why in the optimal solution there is no center opened in their districts. Counterfactual questions as "What is the minimal change in the budget, or in the costs for building a center in their district, or in the population density, that leads to an optimal solution where a center is chosen in their district?"

Of course similar counterfactual explanations are needed in other classical facility problems. For example, one of the authors optimized the physical distribution structure for Philips in Europe. Several distribution centers were closed, and new ones were opened. Of course, the management of these centers that were going to be closed has to be explained why a closure is necessary. Counterfactual explanations are the ideal tool for that.

Network flow problem. Many supply chain problems can be modeled as (multi-commodity) network problems. For example, in the linear optimization model for WFP's food supply chain (Peters et al., 2021, 2022), the diet problem is integrated into a network model to model the multi-modal transportation from the supplier to the beneficiaries. Suppose that a certain port is not used according to the optimal solution of the linear optimization model. The authorities of this port would like to know what is the minimal reduction in costs such that it is used in the optimal solution. Similar counterfactual explanations are needed for potential transportation companies.

We present more examples in Appendix A.

Most of the above examples are at the tactical or even strategical level. Indeed, especially for strategical decisions that affect multi-stakeholders, counterfactual explanations are needed. However, we emphasize that such explanations are also often needed in more operational decisions, where it can affect the personal lives. Two examples taken from the Franz Edelman Award finalists can be found in Appendix A.

In the examples above, multiple stakeholders can be involved, and these stakeholders need counterfactual explanation not only for the sake of explanation, but also for actually changing the input parameters. Counterfactual explanations can thus also be used as a tool for negotiation among multiple stakeholders.

Counterfactual explanations could also be valuable in single stakeholder environments. Minimal changes in input parameters leading to a certain solution being optimal gives much insight in the decision problem. Feasibility ensuing counterfactual explanations are specifically very valuable for answering questions like "What is the minimal change in the input data such that the problem becomes feasible?"

Notice that in all the examples mentioned above, even for the simplest diet problem, the *factual* explanations do not work. The linear optimization model and the simplex or interior point methods are too difficult to explain to a non-expert. However, "what-if scenario analysis" could provide partial explanation in some cases. For example, enforcing in the facility location model that a certain center is opened, one could optimize for the overall accessibility, and then calculate the accessibility decrease. The explanation is then: "If we open this center, then there will be a reduction of x% in overall accessibility. Since in general obtaining explanations by "what-if scenario analysis" is computationally much easier than by counterfactual explanations, we advocate to use both.

We finally point out that the result of counterfactual explanation could also be that the minimal change is extremely small, or even no change has to be performed. The last case could happen when there are multiple optimal solutions and at least one of it already has the desired properties. For those cases, we argue that *factual* explanations should be added, based on secondary criteria not included in the linear optimization model.

1.2. Related work

In Korikov et al. (2021) and Korikov and Beck (2021, 2023) the authors study a definition of counterfactual explanations for integer optimization problems. This definition coincides with our weak CE definition; see Section 2. Korikov et al. (2021) introduce the concept and study it for the case where only the objective function parameters of the problem may be adjusted. Furthermore, they restrict their approach to the case that the desired solution property may only be defined on a single variable. Additionally, they assume that no such desired solution is optimal for the present problem. Korikov and Beck (2021) connect the idea of CEs to inverse optimization and use inverse constraint programming to solve the problem where again only the objective function parameters may be adjusted. Finally, Korikov and Beck (2023) generalize the latter works to the case that the constraints for the desired outcome can be defined on all variables. They develop a constraint generation algorithm which can solve the CEs problem for discrete optimization problems if only the objective parameters may be mutable. To this end a sequence of computationally expensive bilinear mixed-integer problems has to be solved. We relax all the assumptions of the former works and propose a new notion of CEs

which is computationally tractable in the most general case for linear optimization problems.

Counterfactual explanations were also studied in Bogetoft et al. (2024) to explain Data Envelopment Analysis models. Regarding the so-called Farrell input-oriented efficiency, which can be calculated by solving a linear optimization problem, the authors ask the counterfactual question: what is the minimum change in a firms input parameters such that a certain efficiency level is obtained. The resulting problem is a bilevel optimization problem which is in fact a special case of our weak counterfactual explanation problem which we will discuss in this work. Forel et al. (2023) assume that the parameters of the present optimization problem are derived by predictions based on additional context parameters (*e.g.*, the weather, day or temperature). They consider the smart-predict-then-optimize method of Bertsimas and Kallus (2020), where random forests and nearest neighbor predictions are used. They adapt the idea of CEs to calculate CEs in the context parameter space where changes are only considered in the objective parameters of the problem.

A different concept for explaining optimization problems, which is not connected to counterfactual explanations, was developed by Aigner et al. (2024). The authors present a data-driven explanation term, based on historical solutions for the same problem class, which is added to the objective function of the optimization problem to increase the explainability of the solution.

Finally, Goerigk and Hartisch (2023) propose an inherently interpretable model, which provides interpretability of the derived solutions. The authors calculate a decision tree and a small set of solution such that each future problem instance is mapped by the tree to one of the determined solutions. While this approach can be generalized to constraint parameters, it is mainly studied for the objective function parameters.

In the earliest works, Dantzig (1963) develops general linear programming, in which (some of) the parameters of the linear optimization problem are variables too. This method is extensively extended to a much wider class of problems (Gorissen et al., 2022). We will use these techniques to calculate CEs for linear optimization.

There is a strong relationship between CEs and inverse optimization, although they differ in their input–output relationship, specificity of solution, and computational complexity. In inverse optimization, the desired optimal solution is usually a predetermined input, while in finding CEs, the optimal solution is not predefined; only certain constraints that the desired solution must meet are provided. This requirement for the solution to be optimal makes the CE problem significantly more challenging in terms of computational complexity. For a more detailed exploration of inverse optimization, we refer to the recent work by Chan et al. (2025). While there are studies on partial inverse optimization (where some of the decision variables may be variable) these usually allow changes only in the parameters of the objective function — ensuring that the feasible region remains unchanged. The latter problems still result in computationally complex bilinear problems; see, for instance, (Wang, 2013).

Several papers (*e.g.*, Amaral et al., 2008, Barratt et al., 2021, and Moosaei & Hladík, 2021) study the question "What is the minimal change in the input data such that the problem becomes feasible?" This is a simple version of the counterfactual explanation concept. Especially the computationally tractable methods for relative counterfactuals developed in this paper could also be used to answer such feasibility questions.

Parametric optimization and sensitivity analysis are related concepts where changes in single parameters (leading to changes in multiple coefficients in the former case) are studied. However, the goal of these concepts is to analyze the region of changes which can be performed without changing the optimal solution (or optimal basis). Moreover, sensitivity analysis can be seen as factual explanation: it analyzes the effect on the optimal solution when we change the values of the problem parameters.

1.3. Contributions

The main contribution in this work is the development of a new notion of CE for linear optimization, called "relative CE", for which we show that it can be computed in the same order of time as the underlying linear problem. We show that this computational tractability holds for the general case where both objective and constraint parameters may be changed and under mild assumptions. We conduct an extensive computational study on the relative CEs using linear optimization problems from the NETLIB library. The results indeed show that the relative CEs can be calculated in a similar time frame as the underlying linear optimization problems. Moreover, we analyze two other relevant CE definitions related to the ones used in the literature. We present theoretical properties which indicate that these two CEs are much harder to compute than the relative CE which is confirmed by a real-data case study. To ensure reproducibility of our results, we provide a dedicated repository[1] for replicating our experiments.

2. Definitions for counterfactual explanations

Consider an instance of a linear programming (LP) problem of the form

$$\begin{aligned} &\min\ c^\top x \\ &s.t.\quad Ax \geq b, \\ &\qquad x \geq 0, \end{aligned} \tag{1}$$

where $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. This problem can be represented by its corresponding problem parameters (c, A, b). An optimization algorithm for such an LP can be interpreted as a function which maps every instance $(c, A, b) \in \mathcal{H}$ to an optimal solution x^* of the corresponding LP. For a given *factual instance* $(\hat{c}, \hat{A}, \hat{b}) \in \mathcal{H}$, a counterfactual explanation is a –preferably similar– instance $(c, A, b) \in \mathcal{H}$ for which the optimal solution lies in a *favored solution space*, $\mathcal{D}(\hat{x})$. This space does not contain the optimal solution $\hat{x}$ of the factual instance. In other words, a CE is an update in the optimization parameters such that the optimal solution of the LP with the updated parameters has a given list of desired properties. The main concept described above is visualized in Fig. 1.

Note that in general more than one solution can be optimal for an LP and hence the calculated optimal solution may depend on the choice of the optimization algorithm. There can also be solutions, which are only feasible for the LP, but they return smaller objective function values than a fixed target. These observations give rise to extend the current framework for three different types of CEs, namely *relative counterfactual explanations*, *weak counterfactual explanations*, and *strong counterfactual explanations*.

We start with the *present problem* given as

$$\begin{aligned} &\min\ \hat{c}^\top x \\ &s.t.\quad \hat{A}x \geq \hat{b}, \\ &\qquad x \geq 0, \end{aligned} \tag{PP}$$

where $\hat{c} \in \mathbb{R}^n$ and $\hat{A} \in \mathbb{R}^{m\times n}$ and $\hat{b} \in \mathbb{R}^m$. The present problem can be interpreted as the optimization problem which was solved by the *decision maker* to obtain an optimal decision $\hat{x}$. Each variable can be interpreted as being related to what we call a *stakeholder*. In the diet problem example, the decision maker is the organization (including operations research analysts) that has control over the data as well as the optimization problem, and that is responsible for implementing the final decision. Each variable corresponds to a product sold by a food supplier, which is the stakeholder.

Assume that a given subset of problem parameters is mutable and can be changed within a certain feasible region. To this end, we define

[1] https://anonymous.4open.science/r/CE4LOPT-9CD5/

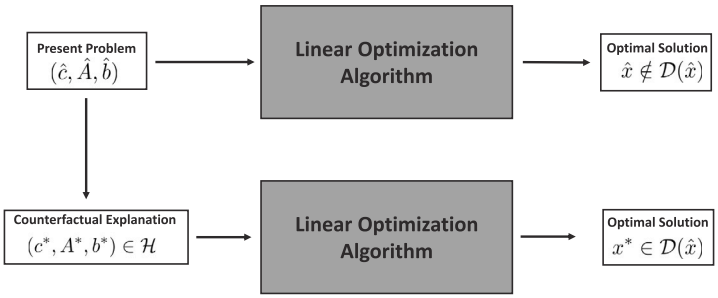

Fig. 1. An overview over the concept of counterfactual explanations for linear optimization problems.

the *mutable parameter space* $\mathcal{H} \subseteq \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^m$ which contains the parameters of the present model, *i.e.*, $(\hat{c}, \hat{A}, \hat{b}) \in \mathcal{H}$. The mutable parameter space can be defined by intervals for each parameter in which the parameter changes or even by a polyhedral or ellipsoidal structure if the allowed changes of parameters depend on each other. For example, if the stakeholder, influenced by the optimization process, has access to change the parameters of the ith column of A (denoted as A_i) and the ith cost parameter c_i, the mutable parameter space is defined as

$$\mathcal{H} = \left\{ (c, A, b) : c_j = \hat{c}_j \text{ and } A_j = \hat{A}_j, \forall j \neq i,\ b = \hat{b},\ c_i \in [\underline{c}_i, \overline{c}_i],\ A_i \in \left[\underline{A}_i, \overline{A}_i\right] \right\},$$

where $\underline{c}_i, \overline{c}_i, \underline{A}_i, \overline{A}_i$ are the corresponding upper and lower bounds on the mutable parameters. In the diet problem the objective parameter c_i corresponds to the price of a product and the column A_i corresponds to the different nutrient values of the product. For every parameter, the stakeholder can define her mutable space depending on how much she is able to change the parameters.

Finally, we define the *favored solution space* $\mathcal{D}(\hat{x})$ which is the set of solutions x that are favored by the stakeholder who is influenced by the decision $\hat{x}$. For example, in the diet problem the favored solution space could be the set of solutions, for which the stakeholder sells at least 5% more of her product i, *i.e.*,

$$\mathcal{D}(\hat{x}) = \left\{ x \geq 0 :\ x_i \geq 1.05\hat{x}_i \right\}.$$

Note that the set $\mathcal{D}(\hat{x})$ does not necessarily has to overlap with the feasible region of the present problem. Furthermore, the favored solution space can be independent of $\hat{x}$ and we denote it as $\mathcal{D}$ in this case.

2.1. Relative counterfactual explanations

The stakeholder, who is influenced by the decision $\hat{x}$, can ask the following counterfactual question:

> "*What is the minimal change of the mutable parameters I have to make such that a solution from the favored solution space changes the objective value at most by a fixed factor?*"

In the following we assume that for the optimal value of the present problem (PP), it holds that $\hat{c}^\top \hat{x} \geq 0$. The latter question leads to the following definition of relative counterfactual explanation.

Definition 1 (*Relative Counterfactual Explanation*)**.** For a given factor $\alpha \in [0, \infty)$ a relative counterfactual explanation is a point $(c, A, b) \in \mathcal{H}$ such that there exists a feasible solution in

$$\left\{ x \geq 0 :\ Ax \geq b,\ c^\top x \leq \alpha \hat{c}^\top \hat{x} \right\} \cap \mathcal{D}(\hat{x}).$$

In contrast to weak and strong CEs (to be defined later), this definition does not require optimality of a solution x in the favored solution space, but requires only feasibility instead. Note that in case the factor α is smaller than one, we are aiming for an improvement of the optimal objective function value, while for $\alpha \geq 1$ a certain deterioration of the objective function value is accepted. See Fig. 2 for a graphical illustration of the relative CE concept.

Note that usually the stakeholder is interested in the smallest changes of the mutable parameters such that the latter definition holds. More formally the stakeholder is looking for a relative CE $(c, A, b) \in \mathcal{H}$ which is closest to the point $(\hat{c}, \hat{A}, \hat{b})$ in a distance metric $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$.

2.2. Weak counterfactual explanations

Often the decision maker implements optimal solutions for a given present problem. Hence, the stakeholder can be interested in parameter changes which lead to certain properties for an optimal solution. The stakeholder, who is influenced by the decision $\hat{x}$, can ask the following counterfactual question.

> "*What is the minimal change of the mutable parameters I have to make such that a solution from the favored solution space is optimal?*"

This leads to the following formal definition.

Definition 2 (*Weak Counterfactual Explanation*)**.** A weak counterfactual explanation is a point $(c, A, b) \in \mathcal{H}$ such that there exists an optimal solution x^* of Problem (1) which lies in the *favored solution space* $\mathcal{D}(\hat{x})$.

In other words, a weak counterfactual explanation is an update of the mutable parameters of the present problem such that at least one optimal solution exists which belongs to the favored solution space. However, since multiple solutions can be optimal, there is no guarantee that the optimal solution implemented by the decision maker is contained in the favored solution space. Note that usually the stakeholder is interested in the smallest changes of the mutable parameters such that the latter definition holds. More formally, the stakeholder is looking for a weak CE $(c, A, b) \in \mathcal{H}$ which is closest to the point $(\hat{c}, \hat{A}, \hat{b})$ in a distance metric $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$. See Fig. 3 for a graphical illustration of the weak CE concept.

In the following proposition we show a property of weak CEs which was already used in Korikov et al. (2021) to define the concept of CEs.

Proposition 3. *The point (c, A, b) is a weak CE if and only if*

$$\begin{array}{lll} \min & c^\top x & \\ s.t. & Ax \geq b & = \\ & x \geq 0 & \end{array} \qquad \begin{array}{ll} \min & c^\top x \\ s.t. & Ax \geq b \\ & x \in \mathcal{D}(\hat{x}) \\ & x \geq 0, \end{array}$$

i.e., the constraints in $\mathcal{D}(\hat{x})$ are redundant.

In fact the equation in the proposition is used in Korikov et al. (2021) to describe counterfactual explanations, which shows that indeed the authors study weak counterfactual explanations. The following proposition shows connections between weak and relative CEs.

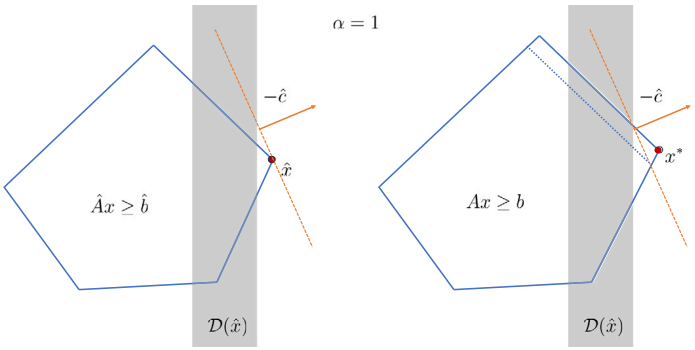

Fig. 2. Graphical illustration of the relative CE concept for $\alpha = 1$, i.e., we are looking for a shift in problem parameters for which a solution in $\mathcal{D}(\hat{x})$ is feasible and has the same costs as the present problem. One bounding hyperplane of the feasible region is slightly shifted to get a feasible solution which lies in $\mathcal{D}(\hat{x})$ and has the same costs as the present solution $\hat{x}$ (i.e., lies on the orange line). On the right, such a solution is the intersection of the orange and the blue line with the gray region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

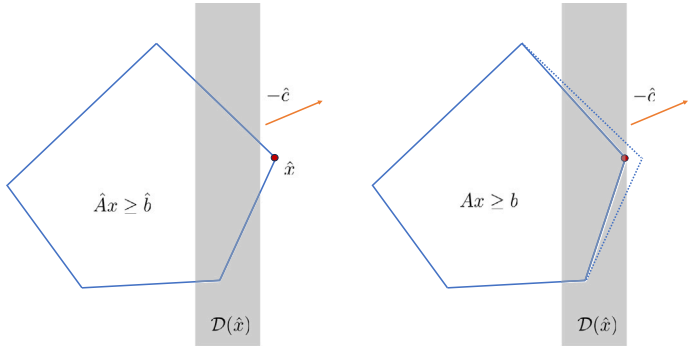

Fig. 3. Graphical illustration of the weak CE concept. Two of the bounding hyperplanes of the feasible region are slightly shifted such that the optimal solution lies in $\mathcal{D}(\hat{x})$ (the gray region).

Proposition 4. *From the definitions of weak and relative CEs, we obtain the following:*

(i) If $\alpha \leq 1$ and the parameters of the present problem $(\hat{c}, \hat{A}, \hat{b})$ are a relative CE, then from the definition it follows that there exists an $x' \in \mathcal{D}(\hat{x})$ such that $\hat{A}x' \geq \hat{b}$ and $\hat{c}^\top x' \leq \hat{c}^\top \hat{x}$. Since $\hat{x}$ is an optimal solution for the present problem, x' must be an optimal solution for the present problem as well and hence $(\hat{c}, \hat{A}, \hat{b})$ is a weak CE as well.

(ii) For every weak CE, there exists an α such that the same point is also a relative CE.

Note that in case (i) of this proposition, since $\hat{x} \notin \mathcal{D}(\hat{x})$, the problem has multiple optimal solutions and at least one of it lies in $\mathcal{D}(\hat{x})$.

2.3. Strong counterfactual explanations

It is important to note that in the definition of weak CEs, we only require that there *exists* an optimal solution of the new optimization problem which fulfills the requirements in $\mathcal{D}(\hat{x})$. However, depending on the optimization algorithm the decision maker uses, this solution may not be implemented, since there might be alternative optimal solutions. To avoid this issue, we define the strong version of counterfactual explanations. Here, the stakeholder which is influenced by the decision $\hat{x}$ asks the following counterfactual question.

> "*What is the minimal change of the mutable parameters I have to make, such that all optimal solutions are contained in the favored solution space?*"

This leads to the following definition.

Definition 5 (*Strong Counterfactual Explanation*)**.** A strong counterfactual explanation is a point $(c, A, b) \in \mathcal{H}$ such that the set of optimal solutions $\mathcal{X}^*$ of problem (1) lies in the *favored solution space* $\mathcal{D}(\hat{x})$, *i.e.*, $\mathcal{X}^* \subset \mathcal{D}(\hat{x})$.

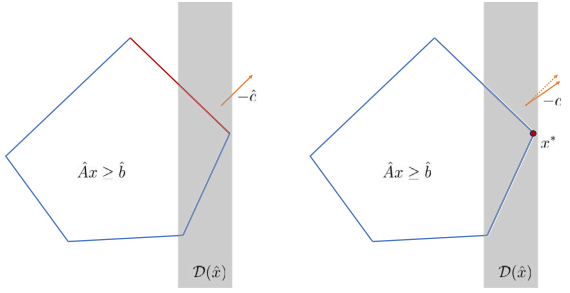

Fig. 4. Graphical illustration of the strong CE concept. Not all optimal solutions (denoted in red) lie in $\mathcal{D}(\hat{x})$ (the gray region). The objective parameters are slightly changed such that only a single solution is optimal and hence all optimal solutions lie in $\mathcal{D}(\hat{x})$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The difference here to the weak version is that we require *all* optimal solutions of the new optimization problem to fulfill the requirements in $\mathcal{D}(\hat{x})$. This is an important difference, since it would guarantee that a decision from the favored solution space will be implemented by the decision maker, independently of the solution method used to determine the optimal decision; see Fig. 4 for a graphical illustration of the strong CE concept.

Note that usually the stakeholder is interested in the smallest changes of the mutable parameters such that the latter definition holds. More formally, the stakeholder is looking for a strong CE $(c, A, b) \in \mathcal{H}$

which is closest to the point $(\hat{c}, \hat{A}, \hat{b})$ in a distance metric $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$.

Proposition 6. *From the definitions of weak and strong CEs it follows:*

1. *Every strong CE is also a weak CE.*
2. *If the set of optimal solutions for a weak counterfactual explanation (c, A, b) is a singleton, then (c, A, b) is a strong CE as well.*

2.4. Practical implications

In this section, we discuss the different ways of how the different counterfactual explanation concepts presented in the previous subsections can be used in practice. One of the main questions in each application is: "Who is providing the counterfactual explanation to whom?"

Decision maker provides explanations to the stakeholder. A frequently occurring situation in practice is that a decision maker who has ownership on the data and the optimization problem is calculating the solution $\hat{x}$ which is afterwards implemented while the stakeholder (*e.g.*, supplier) does not have access to the optimization problem. This is the case in all the examples presented in Section 1.1. For example in the diet problem, the decision maker is calculating an optimal solution of the corresponding linear optimization problem to decide how much of each product is bought. In this case, the supplier asks the decision maker what would be the minimal change in prices she has to perform such that the decision made by the decision maker would be to buy at least a certain amount of a certain product from her. If she asks for a strong CE, then the decision maker will return the minimal change in prices she has to perform such that for any optimal solution (independent of the solution algorithm), the required amount will be purchased from her. In contrast, if she asks for a weak CE, then the decision maker can return the minimal changes in prices together with the information how much of the product would be purchased from her after the change of the prices. In this case it could be that the final decision is not meeting her requirements, since not in every optimal solution the requirements have to be fulfilled for a weak CE. Additionally, if the decision maker changes her solution algorithm (or its settings) in the future, then the final decision may also change. In case of a relative CE, the supplier has to provide the factor α, and the decision maker will return again the minimum change in prices. However, here the decision maker has to decide if an increase/decrease of the costs by a factor of α is desirable, and if a corresponding solution can be implemented.

Analysis of the problem. Especially the concept of relative CEs can be used to perform an analysis of the problem faced by a decision maker. In the facility location problem, to optimize geospatial accessibility to supplies in case of a disaster the decision maker may want to analyze the effects of parameter changes on the rising costs. For example, an increase of the population in a certain region can effect the costs. In case of a what-if analysis, the decision maker can answer questions of the form: if the population in every region increases by 2% what will be my optimal costs? However, by solving the relative CE problem, we can answer questions now of the form: "What is the minimal increase of the population such that my costs will increase by at most 3%?"

Feasibility analysis. The concept of CEs is also beneficial when the present problem is infeasible, and the decision maker wants to find the smallest changes in the parameter space leading to a feasible problem. In this case, the weak CE problem can be solved with $D(\hat{x}) = \mathbb{R}^n$. Consider for example a network flow problem where the capacities are given and the stakeholder defines the demand and supply for each node. It could be that for the given situation there is no feasible flow. The decision maker can now ask the question: "What is the minimum change in the capacity parameters that I have to perform such that there exists a feasible flow?" In this situation, we could also apply relative CEs without the constraint for the objective value. Furthermore, the concept of strong CEs is equivalent to weak CEs in this example.

3. Relative counterfactual explanations

In the following, we assume for the optimal value of the present problem that $\hat{c}^\top \hat{x} \geq 0$ holds. As we will see, relative counterfactual explanations are easier to compute than weak or strong CEs due to the absence of the optimality condition. In fact, the relative counterfactual explanation problem can be formulated as

$$\text{(RCEP)}: \quad \min_{x,c,A,b} \ \delta\left((c, A, b), (\hat{c}, \hat{A}, \hat{b})\right) \tag{2a}$$

$$s.t. \quad c^\top x \leq \alpha \hat{c}^\top \hat{x}, \tag{2b}$$

$$Ax \geq b, \tag{2c}$$

$$x \in D(\hat{x}), \tag{2d}$$

$$(c, A, b) \in \mathcal{H}, \tag{2e}$$

$$x \geq 0. \tag{2f}$$

Note that the constraints contain bilinear terms. While later we will see that the projection onto the x-space is convex, this may be not the case for the projection onto the (c, A, b)-space as the following example illustrates.

Example 7. Let $c = 1$, $A = (\alpha, -\alpha, 1, -1)^T$, $b = (1, -1, \beta, -\beta)^T$, $0 \leq x_1 \leq 1$, $0 \leq \alpha \leq 2$, and $0 \leq \beta \leq 1$. Then the projection of the feasible region of the relative counterfactual problem (RCEP) onto the (α, β) space is

$$\begin{aligned}&\{(\alpha, \beta) \mid \exists x_1 : x_1 = \beta,\ \alpha x_1 = 1,\ 0 \leq x_1 \leq 1,\ 0 \leq \alpha \leq 2,\ 0 \leq \beta \leq 1\}\\ &= \left\{(\alpha, \beta) \mid \beta = \frac{1}{\alpha},\ \frac{1}{2} \leq \alpha \leq 1\right\},\end{aligned}$$

which is not convex.

The latter result already indicates the hardness of the problem in its general version. While we will later define assumptions under which RCEP is tractable, we first prove that the RCEP is strongly NP-hard.

Theorem 8. *The RCEP is strongly NP-hard, even if the mutable parameters only affect the constraints.*

Proof. We reduce the strongly NP-hard 3-partition problem to RCEP. Given a set $\mathcal{E}$ of $3q$ elements where $q \in \mathbb{N}$, a bound $B \in \mathbb{Z}_+$ and weight $w(e)$ for every element $e \in \mathcal{E}$, such that $\frac{B}{4} < w(e) < \frac{B}{2}$ and $\sum_{e \in \mathcal{E}} w(e) = qB$, the 3-partition problem asks if the set $\mathcal{E}$ can be partitioned into q disjoint subsets $E_1, \ldots, E_q \subseteq \mathcal{E}$, such that $\sum_{e \in E_i} w(e) = B$ for all $i \in [q]$. The latter conditions ensure that each subset E_i must contain exactly three elements if the answer to the problem is yes. This problem is known to be strongly NP-hard; see Garey and Johnson (1979).

First, note that the 3-partition problem can be modeled as a binary feasibility problem as follows. Consider variable $x_{ei} \in \{0, 1\}$ for all $e \in \mathcal{E}$ and $i \in [q]$ which obtains value 1 if element e is assigned to subset E_i and 0 otherwise. Then, if the following system of equalities has a solution, the answer to the corresponding 3-partition instances is *yes*, and otherwise *no*:

$$\sum_{i=1}^{q} x_{ei} = 1, \quad \forall e \in \mathcal{E}, \tag{3a}$$

$$\sum_{e \in \mathcal{E}} w(e) x_{ei} = B, \quad \forall i \in [q], \tag{3b}$$

$$x_{ei} \in \{0, 1\}, \quad \forall i \in [q], e \in \mathcal{E}. \tag{3c}$$

The constraints (3a) ensure that each element $e \in \mathcal{E}$ is assigned to exactly one subset E_i. Constraints (3b) ensure that the sum of the weights of the elements assigned to each subset E_i is equal to B. We

now create an instance of the RCEP as follows: The latter equation system can be described in matrix form and we define $W \in \mathbb{R}^{(q+|\mathcal{E}|)\times q|\mathcal{E}|}$ as the constraint matrix and $r \in \mathbb{R}^{q+|\mathcal{E}|}$ as the right-hand-side vector, i.e., the equation system (3a)–(3b) is written as

$$Wx = r,$$

where x is the vector containing the components x_{ei} for all $e \in E$ and $i \in [q]$. Consider the following present problem,

$$\min\ \hat{c}^\top x + 1 \tag{4}$$
$$s.t.\quad \hat{h}_{ei} x_{ei} = \hat{h}_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{5}$$
$$x_{ei} \le \hat{h}_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{6}$$
$$x \ge 0, \tag{7}$$

where $\hat{c}$ is an arbitrary vector and $\hat{h} = \mathbf{0}$. Then Constraints (6) and (7) ensure that $x = 0$ is the only feasible solution, and hence $\hat{x} = \mathbf{0}$ is the present optimal solution. Consider now the following RCEP,

$$\inf_{x,h}\ \delta\left(h, \hat{h}\right) \tag{8a}$$
$$s.t.\quad \hat{c}^\top x + 1 \le M(\hat{c}^\top \hat{x} + 1), \tag{8b}$$
$$h_{ei} x_{ei} = h_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{8c}$$
$$x_{ei} \le h_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{8d}$$
$$x \in \mathcal{D}, \tag{8e}$$
$$h \in \mathcal{H}, \tag{8f}$$

where δ is the ℓ_1-distance, $\mathcal{D} = \{x \ge 0 : Wx = r\}$, $\mathcal{H} = \mathbb{R}^{\mathcal{E}\times[q]}$, $\hat{h} = 0$ and M is a big-M value which makes the corresponding constraint redundant for all feasible x. This big-M value exists, since for every feasible solution it must hold $x_{ei} \in [0,1]$. Note that (8) is an instance of RCEP where the mutable parameters appear only in the constraints.

The constraints (8c)–(8d) can only be feasible if $h_{ei} \ge 1$ (which enforces $x_{ei} = 1$) or if $h_{ei} = 0$ (which enforces $x_{ei} = 0$). Since we do not have any restrictions on h, Constraints (8c)–(8d) allow all binary solutions in $\{0,1\}^{\mathcal{E}\times[q]}$ and especially all feasible solutions x in (8) must be binary.

Since Constraint (8b) is redundant, and (8e) enforces the constraints $Wx = r$, we can conclude that (8) has a feasible solution if and only if the equation system (3) has a solution which is the case if and only if the answer to the 3-partition problems is *yes*. This proves the result.

Note that instead of the 3-partition problem, any other strongly NP-hard problem which can be modeled as a binary feasibility problem could be used in the proof. □

For the rest of this section, we make the following assumptions which we will need to show that the RCEP (in contrast to Example 7) can be transformed into a convex problem. Furthermore, the assumptions ensure that an optimal solution of the RCEP always exists. In the following, we denote $[n] = \{1, 2, \dots, n\}$ for any positive integer n.

Assumption 9. The mutable parameter space $\mathcal{H}$ is compact, convex and *columnwise*, *i.e.*, we have $\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_n \times \mathcal{H}_{n+1}$ where $(c, A, b) \in \mathcal{H}$ if and only if $\left(c_j, A_j\right) \in \mathcal{H}_j$ for all $j \in [n]$ and $b \in \mathcal{H}_{n+1}$ and all sets $\mathcal{H}_1, \dots, \mathcal{H}_{n+1}$ are bounded and convex.

Assumption 10. The favored solution space $\mathcal{D}(\hat{x})$ is a compact and convex set.

Assumption 11. The distance measure δ is continuous and satisfies

$$\delta\left((c, A, b), (\hat{c}, \hat{A}, \hat{b})\right) = \sum_{j=1}^{n} \delta_j\left(\left(c_j, A_j\right), \left(\hat{c}_j, \hat{A}_j\right)\right) + \delta_{n+1}\left(b, \hat{b}\right),$$

where $\delta_j : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \to \mathbb{R}_+$, for each $j \in [n]$, and $\delta_{n+1} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$.

Generality and limitations of assumptions. The convexity and compactness assumption for $\mathcal{H}$ (Assumption 9) and $\mathcal{D}(\hat{x})$ (Assumption 10) are not very restrictive for real-world examples, since usually for $\mathcal{H}$ classical norm-balls of appropriate size or even the whole space can be used. If the resulting CE cannot be achieved due to drastic changes, this still provides information about the problem and that no closer and more realistic CE exists. As our experimental setup shows, many counterfactual questions can be modeled with a polyhedral $\mathcal{D}(\hat{x})$. Additionally, we note that the final optimization problem still must be tractable, and hence, convex sets are an appropriate choice. However, integer sets are not covered by our assumptions which could appear if for example the nutrition values in the diet problem can only be changed to a finite number of values (due to the restriction to a small number of different types of the food). The columnwise property of $\mathcal{H}$ is non-restrictive in any application where the mutable parameters can be independently changed (e.g., demands of different products, nutrition values of different foods, prizes of different commodities). Furthermore, if the goal of the counterfactual explanation is solely to provide the decision maker with insights on the behavior of the problem (i.e, the CE does not have to be provided to a stakeholder) we can always assume that every mutable parameter can be changed independently of the others. However, for certain applications it could happen that dependencies of parameters in different columns have to be modeled to derive realistic CEs for the stakeholder (for example, a stakeholder could have a budget constraint limiting the changes of different food prizes in the diet problem) in which case Assumption 9 does not hold. Finally, the separability of the distance function δ in Assumption 11 can be achieved for all ℓ_p-distances.

First, we show under these assumptions that the feasible region in the (c, A, b)-space is compact. The proof is provided in Appendix B. This result is needed to guarantee the existence of an optimal solution of the RCEP.

Theorem 12. *Under Assumptions 9–11, the projection of the feasible region of RCEP onto the (c, A, b)-space is compact.*

Due to Assumption 11, the objective function of the RCEP is continuous, and hence, by the Weierstrass Theorem and Theorem 12, an optimal solution of the RCEP always exists. This leads to the following corollary.

Corollary 13. *Under Assumptions 9–11, either an optimal solution for the RCEP exists or it is infeasible.*

We will now discuss how to transform the RCEP into an equivalent convex problem. One special setting which directly leads to a convex optimization problem is the following: Assume the only constraints in the set $D(\hat{x})$ are of the form $x_i = \rho_i$, where ρ_i is a fixed value, *i.e.*, the favored solution space requires fixing a subset of variables to given values. If at the same time only the parameters in c and A corresponding to the same columns as the fixed variables are allowed to change, then no bilinear products appear in the formulation RCEP, and hence, the problem is a convex optimization problem when δ is convex.

To transform RCEP into a convex problem in the more general setting, we use the linear transformation from Gorissen et al. (2022). A similar linearization was also used in Rada et al. (2019) (see Theorem 4.1). We will first show that by using the variable transformation

$$w_j = c_j x_j,\ u_{ij} = a_{ij} x_j, \quad \forall i \in [m], \tag{9}$$

we can reformulate RCEP as a problem with convex feasible region. Then, we show that this variable transformation leads to a convex problem for several frequently-used classes of distance measures, δ.

Theorem 14. *The feasible region of RCEP after transformation (9) is convex and given as*

$$\mathcal{P} = \{(x, w, U, b) : \quad \mathbf{1}^\top w \le \alpha \hat{c}^\top \hat{x}, \tag{10a}$$
$$U\mathbf{1} \ge b, \tag{10b}$$
$$x \in D(\hat{x}), \tag{10c}$$
$$(w_j, U_j) \in x_j \mathcal{H}_j, \quad \forall j \in [n], \tag{10d}$$
$$b \in \mathcal{H}_{n+1}, \tag{10e}$$
$$x \ge 0\}, \tag{10f}$$

where $x_j\mathcal{H}_j := \{h' : h' = x_j h,\ h \in \mathcal{H}_j\}$ *for every* $j \in [n]$.

Proof. The proof follows the same idea as in Gorissen et al. (2022). Clearly, constraints (10a)–(10c) and (10f) are equivalent to (2b)–(2d) and (2f) after applying the transformation (9). Furthermore, if $x_j = 0$, then constraint (10d) enforces $w_j = 0$ and $U_j = 0$. If $x_j > 0$, then Constraint (10d) is equivalent to

$$\left(\frac{w_j}{x_j}, \frac{U_j}{x_j}\right) \in \mathcal{H}_j,$$

and hence, constraints (10d) and (10e) are equivalent to (2d) by Assumption 9. Note that $x_j\mathcal{H}_j$ is convex for every fixed $x_j > 0$. Since $\mathcal{H}_j$ is bounded, it holds $x_j\mathcal{H}_j = \{0\}$ if and only if $x_j = 0$, which is again a convex set. Together with Assumption 10, the set $\mathcal{P}$ is convex. This proves the desired result. □

From the latter result, we can directly conclude with the following observation.

Corollary 15. *The projection of the feasible region of (RCEP) onto the x-variables is convex.*

The following example shows that if x is not restricted in sign (*i.e.*, $x \ge 0$ is not part of Problem (PP)), then the transformation is invalid, and the projection onto the x-space is not convex.

Example 16. Let $c = 1$, $A = (\alpha, -\alpha)^T$, $b = (1, -1)^T$, $-1 \le x_1 \le 1$, $-1 \le \alpha \le 1$. Then, the projection of the feasible region of the relative counterfactual problem (RCEP) onto the x space is given by

$$\{x \mid \exists \alpha :\ \alpha x = 1,\ -1 \le x \le 1,\ -1 \le \alpha \le 1\} = \{-1, 1\},$$

which is not convex.

The following example shows that if the columnwise (Assumption 9) does not hold, then the transformation is invalid, and the projection onto the x-space is not convex.

Example 17. Consider the following simple example, with

$$c = 0,\ A = \begin{pmatrix} \alpha & -\alpha \\ -\alpha & \alpha \end{pmatrix},\ b = (1, -1)^T,\ 0 \le x_1, x_2 \le 1,\ -1 \le \alpha \le 1.$$

Then, the projection of the feasible region of the relative counterfactual problem (RCEP) onto the x space is given by

$$\{(x_1, x_2) \mid \exists \alpha :\ \alpha(x_1 - x_2) = 1,\ 0 \le x_1, x_2 \le 1,\ -1 \le \alpha \le 1\} = \{(1, 0), (0, 1)\},$$

which is not convex.

Substituting the variable transformation (9) into the objective function of RCEP leads by Assumption 11 to the objective function

$$\sum_{j=1}^{n} \delta_j \left(\left(\frac{w_j}{x_j}, \frac{U_j}{x_j}\right), (\hat{c}_j, \hat{A}_j)\right) + \delta_{n+1}(b, \hat{b}). \tag{11}$$

There are several relevant cases in which RCEP remains convex after applying the variable transformation (9):

1. Suppose only the parameters in b and one column $j \in [n]$ are allowed to change. Hence, in this case the objective function becomes

   $$\delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b}),$$

   where δ_j and δ_{n+1} are convex functions. Using an epigraph variable z, the objective is to minimize $z + \delta_{n+1}(b, \hat{b})$, under the additional constraint

   $$\delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) \le z.$$

   If we fix the value z, then we can add this constraint to the set of convex constraints that define $\mathcal{H}_j$. Then, RCEP reduces to a convex optimization problem. We can solve RCEP by using binary search over the possible values of z.
2. Suppose the objection function is given by

   $$\max_{j \in [n]} \delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b}).$$

   Again, we use an epigraph variable z for this objective function. Then, the objective changes into minimizing $z + \delta_{n+1}(b, \hat{b})$ with one extra constraint

   $$\delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) \le z$$

   added to $\mathcal{H}_j$ for every $j \in [n]$. As in the previous case, this yields a convex optimization problem for a fixed value of z, and problem RCEP can be solved by using binary search over the value of z.
3. Suppose the objective function is given by

   $$\sum_{j=1}^{n} x_j \delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j)) + \delta_{n+1}(b, \hat{b}). \tag{12}$$

   One motivation for this choice is that changes in the parameters of a certain column are counted less in the penalty function when the corresponding x_j is low. Another situation could be when only the parameters in one column are mutable, and in $D(\hat{x})$ the value of the corresponding x_j is specified. After substitution (9), this objective function becomes

   $$\sum_{j=1}^{n} x_j \delta_j\left(\left(\frac{w_j}{x_j}, \frac{U_j}{x_j}\right), (\hat{c}_j, \hat{A}_j)\right) + \delta_{n+1}(b, \hat{b}),$$

   which is jointly convex in (x, w, U, b). Moreover, if $\delta_j((c_j, A_j), (\hat{c}_j, \hat{A}_j))$, $j \in [n]$, are linearly representable (*e.g.*, ℓ_1- or ℓ_∞-norm) then, RCEP is linearly representable.
4. Suppose the objective function is given by

   $$\sum_{j=1}^{n} \delta_j((c_j x_j, A_j x_j), (\hat{c}_j \hat{x}_j, \hat{A}_j \hat{x}_j)) + \delta_{n+1}(b, \hat{b}).$$

   After substituting (9), this objective function becomes

   $$\sum_{j=1}^{n} \delta_j((w_j, U_j), (\hat{c}_j \hat{x}_j, \hat{A}_j \hat{x}_j)) + \delta_{n+1}(b, \hat{b}),$$

   which is jointly convex in (w, U, b).

4. Optimality-based counterfactual explanations

In this section we study weak and strong CEs. The main difference to the relative CEs is that optimality conditions are required, since by definition a desired solution has to be optimal for the counterfactual explanation. As we will show, this requirement makes the corresponding problems of finding weak and strong CEs computationally much harder. In fact, we will demonstrate that the feasible region of the corresponding problems can have some undesired properties as non-convexity and even disconnectedness.

4.1. Weak counterfactual explanations

The weak counterfactual explanation problem is defined as

$$(\text{WCEP}): \quad \inf_{x,c,A,b} \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \tag{13a}$$

$$s.t. \quad x \in \operatorname*{arg\,min}_{z:Az\geq b,z\geq 0} c^\top z, \tag{13b}$$

$$x \in D(\hat{x}), \tag{13c}$$

$$(c,A,b) \in \mathcal{H}, \tag{13d}$$

where $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$ is a given distance function in the parameter space. Note that in the objective function (13a), we minimize the distance to the parameters of the present problem. Constraint (13b) imposes that each feasible x is an optimal solution for the problem with chosen parameters (c,A,b). Constraint (13c) ensures that the optimal solution x lies in the favored solution space and constraint (13d) implies that we only consider parameter changes inside the mutable parameter space $\mathcal{H}$. The WCEP can be interpreted as an optimistic bilevel problem; see *e.g.*, Dempe et al. (2015), Kleinert et al. (2021) for an overview about bilevel optimization.

The WCEP can be interpreted as an optimistic bilevel optimization problem; see Dempe (2002). It is known that in its general version this class of problems is strongly NP-hard (Marcotte & Savard, 2005). In the following theorem, we prove that also WCEP, as a special case of a bilevel problem, is strongly NP-hard. The proof can be found in Appendix B.

Theorem 18. *The WCEP is strongly NP-hard, even if the mutable parameters only affect the constraints.*

The following theorem shows that the WCEP can be formulated as a bilinear optimization problem. The proof is provided in Appendix B.

Theorem 19. *Problem WCEP is equivalent to the following problem:*

$$(\textit{WCEP'}): \quad \inf_{x,y,c,A,b} \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \tag{14a}$$

$$s.t. \quad c^\top x \leq b^\top y, \tag{14b}$$

$$A^\top y \leq c, \tag{14c}$$

$$Ax \geq b, \tag{14d}$$

$$x \in D(\hat{x}), \tag{14e}$$

$$(c,A,b) \in \mathcal{H}, \tag{14f}$$

$$x, y \geq 0. \tag{14g}$$

The WCEP has the following *structural* properties. Examples for each property are provided in Appendix C.

1. The feasible region of the (c,A,b)-space can be open (see Example 25). Due to this observation we cannot guarantee that an optimal solution of the WCEP exists, which is why we use the infimum instead of the minimum operator for the WCEP.
2. The projection of the feasible region onto the (c,A,b)-space may be non-convex and disconnected, even if $\mathcal{H}$ and $D(\hat{x})$ are convex sets, and even if only the objective parameters in c are allowed to change (see Examples 26 and 27).
3. The projection of the feasible region onto the x-space is convex if the zero-cost vector is contained in $\mathcal{H}$ and $D(\hat{x})$ is convex (see Proposition 28).
4. The projection of the feasible region onto the x-space (and (x,y)-space) can be non-convex (see Example 29).

We conclude that WCEP is in general a hard non-convex problem. If $D(\hat{x})$, $\mathcal{H}$, and $\delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right)$ are representable by linear constraints, then WCEP is a bilinear optimization problem.

4.2. Strong counterfactual explanations

The strong counterfactual explanation problem is defined as follows

$$(\text{SCEP}): \quad \inf_{x,c,A,b} \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right) \tag{15a}$$

$$s.t. \quad x \in D(\hat{x}), \quad \forall\, x \in \operatorname*{arg\,min}_{z:Az\geq b,z\geq 0} c^\top z, \tag{15b}$$

$$(c,A,b) \in \mathcal{H}, \tag{15c}$$

where $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$ is a given distance function in the parameter space. Note that constraint (15b) ensures that all optimal solutions of the problem with selected parameters (c,A,b) are contained in the favored solution space. The SCEP can be interpreted as a certain variant of pessimistic bilevel problems which was studied in Wiesemann et al. (2013).

The SCEP can be interpreted as a pessimistic bilevel optimization problem; see Dempe (2002) for an introduction of pessimistic bilevel optimization. In the following theorem, we prove that also SCEP is strongly NP-hard. The proof can be found in Appendix B.

Theorem 20. *The SCEP is strongly NP-hard, even if the mutable parameters only affect the constraints.*

In the following theorem, we show how to reformulate SCEP as a bilinear problem.

Theorem 21. *Assume the favorable solution space is given by*

$$D(\hat{x}) = \{x \geq 0 : Wx \leq h\}$$

for $W \in \mathbb{R}^{q\times n}$ and $h \in \mathbb{R}^q$. Then, SCEP has the same objective value as the following problem:

$$\begin{aligned}(\textit{SCEP'}): \quad \inf_{c,A,b,\Lambda,\Gamma,\tau} \ & \delta\left((c,A,b),(\hat{c},\hat{A},\hat{b})\right)\\ s.t. \quad & -\Lambda b + \Gamma c \leq h,\\ & c\tau^\top - A^\top\Lambda^\top \geq W^\top,\\ & -b\tau^\top + A\Gamma \geq 0,\\ & (c,A,b) \in \mathcal{H},\\ & \Lambda \in \mathbb{R}_+^{q\times m}, \Gamma \in \mathbb{R}_+^{q\times n}, \tau \in \mathbb{R}_+^q.\end{aligned}$$

The SCEP has the following *structural* properties. Examples for each property are provided in Appendix D.

1. The feasible region of the (c,A,b)-space can be open (see Example 30). Due to this observation we cannot guarantee that an optimal solution of the SCEP exists, which is why we use the infimum instead of the minimum operator for the SCEP.
2. The projection of the feasible region onto the (c,A,b)-space may be non-convex and disconnected, even if $\mathcal{H}$ and $D(\hat{x})$ are convex sets, and even if only the objective parameters in c are allowed to change (see Example 31).

4.3. Testing feasibility

In the following, we provide efficient ways to test if a returned parameter setting (c,A,b) is actually feasible for the three different types of problems:

- To verify that a solution (c,A,b) is a relative CE, one can solve

  $$\min_x \{c^\top x : Ax \geq b, x \in D(\hat{x}),\ x \geq 0\}.$$

  Then, we can check if the optimal value is at most $\alpha\hat{c}^\top\hat{x}$. If this is the case, then the solution (c,A,b) is a relative CE.
- To verify that a solution (c,A,b) is a weak CE, the equality in Proposition 3 can be checked by solving both the standard linear problem and the same problem with the additional constraints in

$D(\hat{x})$. If both problems have the same optimal objective function value, then (c, A, b) is a weak CE.

- There is an efficient preliminary-test that can detect whether solution (c, A, b) is not a strong CE. This can be done by simply solving (1) and checking if the optimal solution fulfills all the requirements in $D(\hat{x})$. If this is not the case, then the solution cannot be a strong CE. To verify that the solution is a strong CE, we have to check whether the following problem is feasible

$$\begin{aligned} &\min_{x} \ c^\top x \\ &s.t. \quad x \in D(\hat{x}), \quad \forall x \in \operatorname*{arg\,min}_{z: Az \geq b, z \geq 0} \ c^\top z. \end{aligned}$$

The robust constraint can again be reformulated into a finite set of constraints by following the steps in the proof of Theorem 21.

5. Numerical experiments

In this section, we present two experiments. First, we present a case study for the diet problem, based on real-world data, and calculate different counterfactual explanations to answer several counterfactual questions. The results indicate that calculating optimal relative CEs can be calculated in the same order of time as the original linear optimization problem while weak or strong CEs are computationally much harder to compute. In the second experiment we calculate relative CEs for all NETLIB instances (NETLIB, 2024) and show that this can be done in about the same time as calculating the optimal solution of the original linear optimization problem.

All algorithms were implemented in Python 3 and executed on a cluster with CPU AMD Rome 7H12 (2x), 64 Cores/Socket, 2.6 GHz and 16 x 16GiB 3200 MHz DDR4 RAM. All optimization problems are solved by Gurobi 10.0.2.

5.1. A case study: The diet problem

In this section, we illustrate the different types of CEs in a practical setting. To this end, we consider the following version of the diet problem, which was studied in Maragno et al. (2025).

We consider a set of different types of food $\mathcal{F}$, containing, *e.g.*, Wheat, Sugar or Beans. Each food type contains a set of nutrients $\mathcal{N}$. The purchasing price per 100 g of food type f is denoted as p_f. For each food type f, the amount of nutrient $v \in \mathcal{N}$ per 100 g is denoted as w_f^v. For each nutrient, the required amount in grams per day is given as req_v. The goal is to decide how much of each food to buy, such that the whole food basket satisfies all nutrient requirements and has minimal costs.

The diet problem is then defined as

$$\begin{aligned} \min \ & \sum_{f \in \mathcal{F}} p_f x_f \\ s.t. \ & \sum_{f \in \mathcal{F}} w_f^v x_f \geq \text{req}_v, \ \forall v \in \mathcal{N}, \\ & x_{\text{Sugar}} = 0.2, \\ & x_{\text{Salt}} = 0.05, \\ & x \geq 0, \end{aligned} \tag{16}$$

where x_{product} denotes the purchased amount of the product (in 100 g). We use the Syria dataset as used in Maragno et al. (2025), where we choose the prices from the Amman region.

An optimal solution of the present problem (16) calculated by Gurobi (with standard settings) is given in Table 1. In the following we calculate strong, weak, and relative CEs for the latter problem, where we study different favored solution spaces and different settings where either only price parameters (*i.e.*, objective parameters) are mutable or both, price parameters and certain nutrition parameters (*i.e.*, objective and constraint parameters) are mutable. We assume that only the nutrition values of the following foods can be changed (*e.g.*, by changing to a different type of product): Milk, Maize, Oil, Wheat, and WSB. We only select a parameter to be mutable if its present value is non-zero.

We implemented formulation (2) to calculate the relative CEs, the formulation from Theorem 19 to calculate weak CEs and the formulation from Theorem 21 to calculate strong CEs. As distance function, we choose the absolute percental deviation

$$\delta\left((p, W), (\hat{p}, \hat{W})\right) = \sum_{f \in \mathcal{F}} \frac{|p_f - \hat{p}_f|}{\hat{p}_f} + \sum_{f \in \mathcal{F}} \sum_{v \in \mathcal{N}} \frac{|w_f^v - \hat{w}_f^v|}{\hat{w}_f^v}.$$

All formulations were implemented in Gurobi with a time limit of 600 s. All calculations for the relative CEs were finished in milliseconds while the calculations for the weak CEs and the strong CEs often hit the time limit. We run Gurobi with standard parameter settings. All solutions in the following tables are the best known solutions found during the time limit. The mutable parameter space $\mathcal{H}$ allows a change of $\pm 50\%$ for every mutable parameter and for the relative CE problems, we always choose $\alpha = 1$.

In the first situation the supplier has supply shortages in Milk and Wheat and cannot provide the requested amounts from the optimal solution. To avoid canceling the delivery, she asks the following counterfactual question.

Counterfactual Question 22. *How much do I have to change my prices (and nutrition values) such that no Milk and no Wheat is requested by the decision maker?*

In this case the favored solution space is $D = \{x \geq 0 : x_{\text{Milk}} = 0, x_{\text{Wheat}} = 0\}$. The corresponding CEs can be found in Table 2.

The results in Table 2 show that, as expected, the percental value of required changes increases from relative to weak and from weak to strong CEs. Note that the relative CE shows the cost-change the supplier has to perform, such that a food basket exists without Milk and Wheat which has the same costs as the current optimal basket. This basket does not have to be optimal after the corresponding cost change. For the weak CE, the price for Wheat is increased (to avoid using Wheat in the optimal solution) and the price for WSB is decreased, leading to WSB in the optimal solution instead of Wheat and Milk. To ensure that every optimal solution does not contain Milk and Wheat, the price changes for the strong CE are slightly larger. If we allow changes in the nutrition values the relative CE remains the same. For the weak CE no substantial changes can be observed; however, the model terminated with an optimality gap of 42.2%. For the strong CE one nutrition value is changed, and no informative optimality gap was reached (due to a lower bound of 0).

In the second situation, the supplier faces an excessive stock level of Maize, and hence, wants to boost her sales. She asks the following counterfactual question.

Counterfactual Question 23. *How much do I have to change my prices (and nutrition values) such that the requested amount of Maize from the decision maker increases by 50%?*

In this case, the favored solution space is $D = \{x \geq 0 : x_{\text{Maize}} \geq 2.04\}$. The corresponding CEs can be found in Table 3.

The results in Table 3 indicate for the case where only price parameters are mutable that the weak and the strong CE are very close to each other. The strong CE is only perturbed by a small value (vanished by rounding to one digit) probably to ensure that only a single solution is optimal instead of a whole face of the polyhedron. Interestingly, while the weak CE calculations hit the time limit, the strong CE could be solved to optimality. In contrast, for the case where nutrition parameters are also mutable, the strong CE calculations ended with an optimality gap of 100%, while for the weak CE, the gap could be improved to 79.8%. Note for all CE types that no constraint parameters are changed.

Table 1
Optimal solution of the diet problem.

Prod. 1	Prod. 2	Prod. 3	Prod. 4	Prod. 5	Prod. 6	Prod. 7
Milk (50.5 g)	Salt (5 g)	Maize (135.8 g)	Sugar (20 g)	Oil (20.6 g)	Wheat (277.1 g)	WSB (72 g)

Table 2
Best known counterfactual explanations after 600 s time limit for Counterfactual Question 22.

Mutable Parameters	Relative CE		Weak CE		Strong CE	
	CE	t in s (opt. gap)	CE	t in s (opt. gap)	CE	t in s (opt. gap)
$p_f f \in \mathcal{F}$	p_{WSB} :↓ 22.4%	0.2	p_{Wheat} :↑ 19.3% p_{WSB} :↓ 18.3%	2.8	p_{Wheat} :↑ 19.6% p_{WSB} :↓ 18.3%	566.0
$p_f f \in \mathcal{F}$ $w^v_{\text{Milk}}, w^v_{\text{Maize}}$ $w^v_{\text{Oil}}, w^v_{\text{Wheat}}, w^v_{\text{WSB}}$	p_{WSB} :↓ 22.4%	1.1	p_{Wheat} :↑ 19.3% p_{WSB} :↓ 18.3%	600.0 (42.2%)	p_{Maize} :↑ 3.9% p_{Wheat} :↑ 14.2% RiboflavinB2(mg) in Milk : ↓ 25.6%	600.0 (100%)

Table 3
Best known counterfactual explanations after 600 s time limit for Counterfactual Question 23.

Mutable Parameters	Relative CE		Weak CE		Strong CE	
	CE	t in s (opt. gap)	CE	t in s (opt. gap)	CE	t in s (opt. gap)
$p_f f \in \mathcal{F}$	p_{WSB} : ↓ 9.9%	0.1	p_{Maize} :↓ 1.1% p_{Wheat} :↑ 26.2%	600.0 (0.82%)	p_{Maize} :↓ 1.1% p_{Wheat} :↑ 26.2%	105.3
$p_f f \in \mathcal{F}$ $w^v_{\text{Milk}}, w^v_{\text{Maize}}$ $w^v_{\text{Oil}}, w^v_{\text{Wheat}}, w^v_{\text{WSB}}$	p_{WSB} :↓ 9.9%	0.5	p_{Maize} : ↓ 1.1% p_{Wheat} :↑ 26.2%	600.0 (79.8%)	p_{Beans} :↓ 30.6% p_{Maize} :↓ 2.6%	600.0 (100%)

Table 4
Best known counterfactual explanations after 600 s time limit for Counterfactual Question 24.

Mutable Parameters	Relative CE		Weak CE		Strong CE	
	CE	t in s (opt. gap)	CE	t in s (opt. gap)	CE	t in s (opt. gap)
$p_f f \in \mathcal{F}$	p_{Lentils} :↓ 41.9%	1.4	p_{Lentils} :↓ 50.0% p_{Wheat} :↑ 25.2%	600.0 (11.4%)	p_{Lentils} :↓ 50.0% p_{Wheat} :↑ 25.2%	600.0 (100%)
$p_f f \in \mathcal{F}$ $w^v_{\text{Milk}}, w^v_{\text{Maize}}$ $w^v_{\text{Oil}}, w^v_{\text{Wheat}}, w^v_{\text{WSB}}$	p_{Lentils} :↓ 41.9%	4.52	p_{Lentils} :↓ 24.9% RiboflavinB2(mg) in Milk : ↓ 28.9% RiboflavinB2(mg) in WSB : ↓ 1.7%	600.0 (68.1%)	p_{Milk} :↑ 2.2% p_{Lentils} :↓ 16.4% p_{WSB} :↑ 30.8% Fat(g) in Oil : ↑ 18.6% RiboflavinB2(mg) in Milk : ↓ 49.6% RiboflavinB2(mg) in WSB : ↓ 5.0%	600.0 (100%)

In the third situation, imagine the decision maker realizes that all the products of the optimal food basket are the ones which are not available on the local market and have to be shipped over larger distances. The decision maker wants to know what the reason for this is and asks the following counterfactual question.

Counterfactual Question 24. *How much do the prices (and nutrition values) have to change such that at least 200g of the food basket are products from the local market?*

We assume that the local products are Lentils, Rice and Chickpeas. In this case the favored solution space is $D = \{x \geq 0 : x_{\text{Lentils}} + x_{\text{Rice}} + x_{\text{Chickpeas}} \geq 2\}$. The corresponding CEs can be found in Table 4.

The results in Table 4 show again that the weak and the strong CEs are very close to each other (the small difference vanished after rounding to one digit) when only the price parameters are mutable. Again, this is probably the case because for the weak CE, multiple optimal solutions exist, and to obtain a strong CE, the objective function has to be slightly changed such that only one solution is optimal that lies in the favored solution space. Interestingly, although the optimality gap is still 100% for the strong CE, the comparison with the weak CE indicates that the optimization model already found the (close to) optimal solution but could not prove optimality during the time limit. In contrast, the optimality gap for the case where also constraint parameters are mutable could be decreased to 68.1% for weak CEs. For the relative CE, similar as for Counterfactual Counterfactual Question 23, the CE is the same for both mutable parameter setups and can be calculated in seconds.

Table 5
Categorization of NETLIB instances.

Type	Category	Intervals
# Variables	Small	$0 \leq n \leq 534$
# Variables	Medium	$534 \leq n \leq 2167$
# Variables	Large	$2167 \leq n \leq 22275$
# Constraints	Small	$0 \leq m \leq 351$
# Constraints	Medium	$351 \leq m \leq 906$
# Constraints	Large	$906 \leq m \leq 16675$

In several situations we found stark numerical instabilities in our experiments when solving the bilinear formulations for weak and strong CEs. These instabilities are probably caused by the possibility that the feasible region of the counterfactual problems can be open as shown in Sections 4.1 and 4.2. Changing the accuracy parameters of Gurobi, or switching off the presolve procedure, could lead to significantly different solutions, and sometimes, it could not be verified by applying

Table 6
Solution time information for the considered classes of NETLIB instances. We show from left to right the following values: the size of the dimension; the size of the number of constraints; the number of random columns for mutable parameter selection; the solution time t in seconds (averaged over all instances which are not infeasible); the time t_{inf} in seconds which was needed to detect infeasibility (averaged over all instances which are infeasible); the time t in seconds to solve the present problem (PP); the time t_{setup} to set up the problem formulation for the present problem.

			Linear Reformulation CE Problem (10)		Present Problem (PP)	
n	m	# mut. col.	t	t_{inf}	t	t_{setup}
Small	Small	1	0.14	0.13	0.13	0.12
		5	0.15	0.13		
		10	0.17	0.15		
	Medium	1	0.31	0.30	0.32	0.30
		5	0.33	0.31		
		10	0.36	0.33		
Medium	Small	1	0.53	0.77	0.53	0.49
		5	0.54	nan		
		10	0.58	nan		
	Medium	1	0.50	0.44	0.50	0.45
		5	0.49	0.44		
		10	0.51	0.45		
	Large	1	1.04	1.03	1.14	1.06
		5	1.05	1.01		
		10	1.08	1.04		
Large	Small	1	11.10	8.27	8.22	7.98
		5	10.50	8.28		
		10	11.08	7.19		
	Medium	1	0.90	1.36	1.34	1.24
		5	0.92	1.35		
		10	0.95	1.38		
	Large	1	4.90	5.36	17.35	3.18
		5	4.42	5.40		
		10	4.95	4.90		

the tests in Section 4.3 that the solution is indeed a weak or strong CE. However, all presented results are for situations without the latter complications.

5.2. NETLIB instances

In this section, we perform experiments for relative CEs on the NETLIB instances (NETLIB, 2024). To this end, we calculate CEs by the convex reformulation (10) with objective (12) using δ_j as the ℓ_1-norm. Using the resulting objective function, we finally obtain a linear optimization problem even for an arbitrary number of columns which may contain mutable parameters.

For every NETLIB instance, we calculate the optimal solution $\hat{x}$ and generate the favored solution space $\mathcal{D}(\hat{x})$ as follows: for every instance, we select three random columns j. If there is no conflict with the variable bounds of the NETLIB instance the favored constraint $x_j \geq 1.05\hat{x}_j$ is added (otherwise $x_j \leq 0.95\hat{x}_j$). If $\hat{x}_j = 0$, then the favored constraint $x_j \geq 0.05$ is added only if there is no conflict with the original variable bounds (otherwise $x_j \leq -0.05$). To generate the mutable parameter space $\mathcal{H}$, we do the following. We draw one, five, and 10 random columns, where we only consider columns for which the original lower bound on the variable in the present problem is non-negative. In each of the selected columns, all parameters are defined as mutable when rounding to zero digits changes the value. For a small set of instances, all problem parameters are integer, hence the latter procedure does not provide any mutable parameters. For this set of instances, all parameters are considered as mutable whose absolute value is larger than 10 and is not a multiple of 10. If we draw five columns, then we combine these with the column from the iteration where we only draw one column. When we draw 10 columns, we combine these with the columns from the iterations where we draw one column and five columns. By this procedure we ensure that the optimal value for a larger number of columns is at least as good as for a smaller number, since all columns from the previous iterations are contained. Every mutable parameter is allowed to change by ±100%.

The optimality factor is set to $\alpha = 1$, *i.e.*, we are looking for the smallest change in the problem parameters such that a solution with the desired properties in $\mathcal{D}(\hat{x})$ exists, and it has the objective function value at least as good as the present problem. For every NETLIB instance and for every number of columns in $\{1, 5, 10\}$, we generate 20 CE instances as described above. We divide the NETLIB instances into nine categories where we classify the dimension n of the problem (*i.e.*, the number of variables) as small if n is at most the 35% quantile, as medium if n is between the 35% and the 70% quantile, and as large if n is above the 70% quantile of the list of dimensions of all NETLIB instances. We classify the number of constraints m of each problem by the same procedure. The detailed information for the categorization is shown in Table 5, and the information on the instance setup is provided in Table E.7 of Appendix E.

Table 6 shows results on the runtime of calculating relative CEs compared to the runtime of the original linear optimization problem. We also show the time to solve the present problem (PP) and the time to set up the problem formulation for the present problem. The latter metric is motivated by our observation that setting up the constraint matrix (although using the sparse-matrix datatype provided by Numpy) takes a significant amount of the runtime for both, the present problem and the counterfactual problems. The results show that the runtime of the present problem is of the same order or sometimes even larger order compared to the runtime of solving the linear relative CE problem. The larger runtime comes from the fact that not all of the relative CE problems are feasible. In Appendix E we compare these results with the bilinear reformulation (2), and show indeed that the linear reformulation can be solved in much shorter time.

6. Conclusion

In this work, we argue that counterfactual explanations constitute a useful tool to provide explainability for linear optimization problems. We present three different types of counterfactual explanations which cover many relevant situations in practical applications. In contrast to weak CEs, the concept of strong CEs considers the case that more than one optimal solution exists and enforces that all of them fulfill the desired conditions. On the other hand relative CEs provide changes in the problem parameters for which a desired solution would not lead to a large increase in objective value. Our theoretical analysis shows that the relative CE problem can be reformulated as a convex problem under mild assumptions. Exploiting this hidden convexity leads to computationally tractable solution methods as our experiments also confirm.

Since the development of counterfactual explanations in optimization is still in its infancy, there are many open research questions which should be studied in future works. While we show negative structural results for the weak and the strong CE problems, it may be possible to detect special cases for which the corresponding problems are computationally tractable. Furthermore, the concepts we developed in this work could be extended to non-linear optimization problems, or to problems involving integer decisions. While for the latter class of problems some works exist, these do not cover the most general case of mutable constraint parameters. Finally, additional desired properties for counterfactual explanations (e.g., sparsity, actionability, coherence, etc.) could be modeled into our problem formulations in a similar fashion as in Maragno et al. (2022).

While we developed the concept of strong CEs to tackle the issue of possible multiple optimal solutions, our experiments show that the changes in problem parameters which have to be performed for a strong CE can be significant or in some situations a strong CE even does not exist. This is undesirable in a practical setting. The reason for developing this concept stems from the fact that we do not consider which solution algorithm is used by the decision maker to derive an optimal solution. A future direction could be to calculate CEs for specific types of optimization algorithms, *e.g.*, the simplex method or the branch-and-bound method. Although this is a challenging task, it would provide a less conservative way to calculate CEs in the strong fashion.

CRediT authorship contribution statement

Jannis Kurtz: Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Conceptualization. **Ş. İlker Birbil:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Dick den Hertog:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization.

Appendix A. Motivating examples

Flood safety problem. Optimization has been used to determine new safety standards for the dike heights in the Netherlands (Brekelmans et al., 2012; Eijgenraam et al., 2017). For each of the 53 so-called dike-ring areas, *i.e.*, an area that is protected by dikes, finally one out of five different safety levels has been chosen. For the management of a specific dike-ring area it is crucial to know what minimal change in the dike-ring area characteristics (as for example the number of people or the economic value in that area) would have led to a higher safety level.

Scheduling trains in the Netherlands. In December 2006, Netherlands Railways introduced a completely new timetable by using sophisticated operations research techniques (Kroon et al., 2009). The first versions of the schedule led to much social unrest for the employees. The schedules were considered as "boring", since for each employee it was more or less the same every day. After these aspects were included in the approach, and the resulting schedule was better explained, the employees finally accepted the new schedule.

Boston public school transport. Bertsimas et al. (2020) describe that optimization methods were developed for Boston Public Schools (BPS) to create a better way to construct bus routes. The goals were improving efficiency, deepening the ability to model policy changes, and realigning school start times. Using this methodology, BPS proposed a solution that would have saved an additional $12 million annually and also shifted students to more developmentally appropriate school start times (*e.g.*, by reducing the number of high school students starting before 8:00 a.m. from 74% to 6% and the average number of elementary school students dismissed after 4:00 p.m. from 33% to 15%). However, 85% of the schools' start times would have been changed, with a median change of one hour. This magnitude of change led to strong vocal opposition from some school communities that would have been affected negatively. Therefore, BPS did not implement the plan.

Appendix B. Proof of theorems

B.1. Proof of Theorem 12

To prove the result we show that every convergent sequence of feasible points in the (c, A, b)-space has a limit which is also feasible. W.l.o.g. we may assume that the objective parameters c are not mutable, since we can always shift them into the constraints by using an epigraph reformulation. Furthermore, w.l.o.g we may assume that the righ-hand-side parameters b are not mutable since we can introduce a new variable x_{n+1} and rewrite the constraint system as

Assume we have an infinite converging sequence $\{A^t\}_{t\in\mathbb{N}}$ of constraint matrices for which RCEP is feasible. We denote the limit of the sequence as $\overline{A}$. This limit lies in $\mathcal{H}$ since it is closed. Then for every t there exists a point $x^t \in \mathcal{D}(\hat{x})$ for which $A^t x^t \geq b$. Since $\mathcal{D}(\hat{x})$ is compact, there exists a converging subsequence $\{x^t\}_{t\in I}$ with limit $\tilde{x} \in \mathcal{D}(\hat{x})$. It holds

$$\overline{A}\tilde{x} = \lim_{t\to\infty, t\in I} A^t \lim_{t\to\infty, t\in I} x^t = \lim_{t\to\infty, t\in I} A^t x^t \geq b,$$

where the latter inequality holds since $A^t x^t \geq b$ for all t and the set $\{v : v \geq b\}$ is closed. By the same argumentation $\tilde{x}$ fulfills all other constraints of RCEP and hence $\overline{A}$ is feasible for RCEP which proves the closedness of the feasible region. Since $\mathcal{H}$ is bounded by assumption this proves compactness. □

B.2. Proof of Theorem 18

We reduce the strongly NP-hard 3-partition problem to WCEP which was already defined in the proof of Theorem 8. We use again the notation $Wx = r$ to model the Eqs. (3a) and (3b). Consider the following present problem,

$$\min\ 0 \tag{B.1}$$

$$s.t.\quad \hat{h}_{ei} x_{ei} = \hat{h}_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{B.2}$$

$$x_{ei} \leq \hat{h}_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{B.3}$$

$$x \geq 0, \tag{B.4}$$

where $\hat{h} = \mathbf{0}$. Then Constraints (B.3) and (B.4) ensure that $x = 0$ is the only feasible solution, and hence $\hat{x} = \mathbf{0}$ is the present optimal solution.

Consider the following WCEP,

$$\inf_{x,h}\ \delta\left(h, \hat{h}\right) \tag{B.5a}$$

$$s.t.\quad x \in S(h), \tag{B.5b}$$

$$x \in D, \tag{B.5c}$$

$$h \in \mathcal{H}, \tag{B.5d}$$

where δ is the ℓ_1-distance, $D = \{x \geq 0 : Wx = r\}$, $\mathcal{H} = \mathbb{R}^{\mathcal{E}\times[q]}$, and

$$S(h) = \arg\min_x 0 \tag{B.6a}$$

$$s.t. \quad h_{ei}x_{ei} = h_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{B.6b}$$

$$x_{ei} \leq h_{ei}, \quad \forall i \in [q], e \in \mathcal{E}, \tag{B.6c}$$

$$x \geq 0. \tag{B.6d}$$

Note that (B.5) is an instance of WCEP where the mutable parameters appear only in the constraints.

First, note that the constraints (B.6b)–(B.6c) can only be feasible if $h_{ei} \geq 1$ (which enforces $x_{ei} = 1$) or if $h_{ei} = 0$ (which enforces $x_{ei} = 0$). Since we are minimizing the distance $\delta\left(h, \hat{h}\right)$ and $\hat{h} = 0$, in any optimal solution h_{ei} must be either 0 or 1 and for any $x \in S(h)$ it holds $x = h$.

We can show now that the system (3) has a feasible solution if and only if the optimal value of (B.5) is lower or equal to $|\mathcal{E}|$. Since (3) is feasible if and only if the answer for the 3-partition problem is *yes*, this proves the result.

First, let x' be any feasible solution of (3). Then, $x' \in D$ and for $h^* = x'$ we have $x' \in S(h^*)$. Hence, h^* is feasible for (B.5). Furthermore, since x' fulfills the Eqs. (3a) it follows that

$$\delta\left(h^*, \hat{h}\right) = \sum_{e\in\mathcal{E}} \sum_{i\in[q]} h^*_{ei} = |\mathcal{E}|,$$

and therefore the optimal value of (B.5) is at most $|\mathcal{E}|$. For the other implication, assume that the optimal value of (B.5) is at most $|\mathcal{E}|$. Especially, this means that (B.5) is feasible and there exists an $\bar{h} \in \mathcal{H}$ for which the Subproblem (B.6) is feasible. As noted above any feasible solution $\bar{x}$ of this subproblem must be binary, and since $\bar{x} \in D$ must hold, there exists a binary $\bar{x}$ with $W\bar{x} = r$. It follows that this $\bar{x}$ fulfills the Eqs. (3a).

Note that instead of the 3-partition problem, any other strongly NP-hard problem which can be modeled as a binary feasibility problem could be used in the latter proof. □

B.3. Proof of Theorem 19

We replace the optimality constraint (13b) by constraints (14b)–(14d) and (14g), where (14d) and (14g) ensure primal feasibility for x, (14c) and (14g) ensure dual feasibility for y, and (14b) ensures that primal and dual objective function values are equal. By the classical strong duality result, it follows that any feasible x must be optimal for the minimization problem in (13b). □

B.4. Proof of Theorem 20

Proof. We perform exactly the same reduction as in the proof of Theorem 18 but use the strong variant of the CE problem. Note, that for any feasible $h \in \mathcal{H}$, the feasible solution x of the subproblem (B.6) is binary and unique, i.e., only one optimal solution exists. Hence, the whole reduction is valid for the SCEP as well. □

B.5. Proof of Theorem 21

First, we can reformulate SCEP as

$$\inf_{c,A,b} \delta\left((c, A, b), \left(\hat{c}, \hat{A}, \hat{b}\right)\right)$$
$$s.t. \quad x \in D(\hat{x}), \quad \forall\, x \in \mathcal{U},$$
$$(c, A, b) \in \mathcal{H},$$

where $U = \{(x, y) \geq 0 \mid c^\top x \leq b^\top y, Ax \geq b, A^\top y \leq c\}$. In the description of U, we use the same optimality conditions as in the proof of Theorem 19. The latter problem is a robust optimization problem with decision dependent uncertainty set. By using the classical constraint-wise dualization trick from robust optimization, we can reformulate the problem as follows: First rewrite the problem as

$$\inf_{c,A,b} \delta\left((c, A, b), \left(\hat{c}, \hat{A}, \hat{b}\right)\right)$$
$$s.t. \quad \max_{x\in\mathcal{U}} w_i^\top x \leq h, \quad i =, 1 \ldots, p,$$
$$(c, A, b) \in \mathcal{H},$$

where w_i is the ith row of matrix W. By using duality theory for linear optimization problems we obtain

$$\max_{x\in\mathcal{U}} w_i^\top x = \min_{\substack{\tau_i, \Gamma_i, \Lambda_i \geq 0,\\ \tau_i c - A^\top \Lambda_i \geq w_i,\\ -\tau_i b + A\Gamma_i \geq 0,}} c^\top \Gamma_i - b^\top \Lambda_i,$$

where τ_i is the dual variable for constraint $c^\top x - b^\top y \leq 0$, Γ_i are the dual variables for the constraints $A^\top y \leq c$ and $\Lambda_i \geq 0$ are the dual variables for constraints $-Ax \leq -b$. When replacing the maximization problem in Constraint i with the dual minimization problem, we can drop the minimum operator, which leads to the formulation

$$\inf_{c,A,b} \delta\left((c, A, b), \left(\hat{c}, \hat{A}, \hat{b}\right)\right)$$
$$s.t. \quad c^\top \Gamma_i - b^\top \Lambda_i \leq h, \quad i =, 1 \ldots, p,$$
$$\tau_i c - A^\top \Lambda_i \geq w_i, \quad i =, 1 \ldots, p,$$
$$-\tau_i b + A\Gamma_i \geq 0, \quad i =, 1 \ldots, p,$$
$$\tau_i, \Gamma_i, \Lambda_i \geq 0, \quad i =, 1 \ldots, p,$$
$$(c, A, b) \in \mathcal{H}.$$

The latter problem can be reformulated in matrix notation as stated in the theorem, with $\tau = (\tau_1, \ldots, \tau_p)$, $\Gamma = (\Gamma_1, \ldots, \Gamma_p)$, and $\Lambda = (\Lambda_1, \ldots, \Lambda_p)$.

We can now dualize each of the maximization problems appearing on the left-hand-side separately (introducing a copy of the dual variables for each one) which leads to the formulation presented in the theorem. □

Appendix C. Examples for weak counterfactual explanations

We start with showing that the feasible region of the (c, A, b)-space can be open.

Example 25. Consider the present problem

$$\min\{x : \ \hat{a}x = \hat{a}, \ x \geq 0\},$$

where $\hat{a} = 0$. The unique optimal solution is $\hat{x} = 0$. Assume the favored solution space $\mathcal{D}(\hat{x}) = \{x \in \mathbb{R} : x \geq 1\}$ and $\mathcal{H} = \left[0, 1\right]$. For every $a > 0$ the optimal solution of the problem is $x = 1$, which lies in $\mathcal{D}(\hat{x})$. Hence, the feasible region for parameter a in WCEP is $(0, 1]$ which is open. In this case the WCEP does not have an optimal solution.

Example 26. The following example shows that the projection of the feasible region of WCEP onto the variable space (c, A, b) may be non-convex; see Fig. C.5 for an illustration. Consider the present problem

$$\max\{x_1 : \ |x_1| + |x_2| \leq 1\},$$

which can be reformulated into a linear optimization problem. Note that we do not require here that $x \geq 0$ as we do in Theorem 19. However, this can be easily achieved by shifting the feasible region into the positive orthant, *e.g.*, by replacing x_i by $x_i - 1$ everywhere in the problem. The problem has the unique optimal solution $\hat{x} = (1, 0)$. Suppose now that the favored solution space is given by

$$D = \left\{x \in \mathbb{R}^2 : \ -0.5 \leq x_1 \leq 0.5\right\}.$$

Furthermore, we assume that only the two objective parameters c_1, c_2 can be changed to any value. Let $\text{int}(S)$ denote the interior of a set S. Then, the following statements hold:

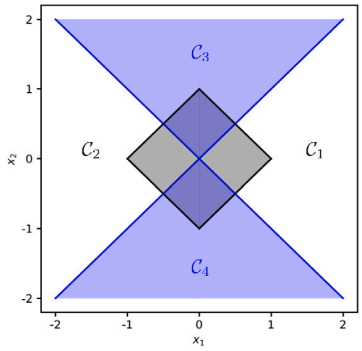

Fig. C.5. The feasible region of the present problem of Example 26 (in gray) and the feasible region of the weak CEs for the objective parameters (in blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- For any $(c_1, c_2) \in \text{int}(C_1)$ where $C_1 = \left\{(c_1, c_2) = \lambda_1 \binom{1}{1} + \lambda_2 \binom{1}{-1}, \lambda_1, \lambda_2 \geq 0\right\}$ the unique optimal solution of the present problem is $x^1 = (1, 0) \notin D$, and hence, C_1 does not contain any feasible weak CE.
- For any $(c_1, c_2) \in \text{int}(C_2)$ where $C_2 = \left\{(c_1, c_2) = \lambda_1 \binom{-1}{1} + \lambda_2 \binom{-1}{-1}, \lambda_1, \lambda_2 \geq 0\right\}$ the unique optimal solution of the present problem is $x^2 = (-1, 0) \notin D$ and hence C_2 does not contain any feasible weak CE.
- For any $(c_1, c_2) \in C_3$ where $C_3 = \left\{(c_1, c_2) = \lambda_1 \binom{1}{1} + \lambda_2 \binom{-1}{1}, \lambda_1, \lambda_2 \geq 0\right\}$ the point $x^3 = (0, 1) \in D$ is optimal, and hence, all points in C_3 are weak CEs.
- For any $(c_1, c_2) \in C_4$ where $C_4 = \left\{(c_1, c_2) = \lambda_1 \binom{1}{-1} + \lambda_2 \binom{-1}{-1}, \lambda_1, \lambda_2 \geq 0\right\}$ the point $x^4 = (0, -1) \in D$ is optimal, and hence, all points in C_4 are weak CEs.

Note that int $(C_1) \cup$ int $(C_2) \cup C_3 \cup C_4 = \mathbb{R}^2$. Thus, the region of feasible weak counterfactuals is $C_3 \cup C_4$, which is non-convex.

The following example shows that the projection onto the feasible region of the variables c, A, b may be disconnected.

Example 27. Consider the same setup as in Example 26 and additionally define the mutable parameter space $\mathcal{H} = \{(c_1, c_2) : 0.5 \leq c_1, c_2 \leq 1.5\}$. Then the feasible region of the weak CEs is given as $(C_3 \cup C_4) \cap \mathcal{H}$ which is a disconnected set.

Proposition 28. *If $(0, \hat{A}, \hat{b}) \in \mathcal{H}$, then the projection of the feasible region of Problem WCEP onto the x-space is convex.*

Proof. Proof For the parameters $(0, \hat{A}, \hat{b})$ all feasible points of the present problem are optimal. Hence, the projection onto the x-space is the feasible set of the present problem intersected with $D(\hat{x})$, which is convex. □

Example 29. The following example shows that the projection of the feasible region of WCEP onto the x-space may be non-convex. Consider the linear optimization problem

$$\begin{aligned} \max\ & c_1 x_1 + c_2 x_2 \\ s.t.\ & x_1 + x_2 \leq 2, \\ & x_1 - x_2 \leq 0, \\ & x_1, x_2 \geq 0, \end{aligned}$$

where the feasible region is shown in Fig. C.6. We assume that only the cost parameters can change, where $(c_1, c_2) \in \{1\} \times [-1, 1]$ and set $D = \mathbb{R}^2_+$. Note that the point $z_1 = (0, 2)$ is feasible and optimal

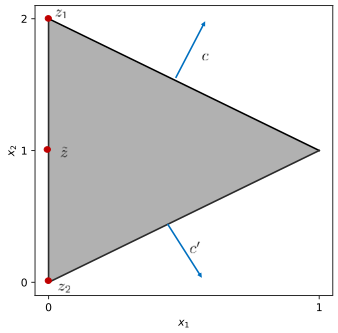

Fig. C.6. The feasible region of the problem of Example 29 (in gray).

for cost vector $c = (1, 1)$. On the other hand, the point $z_2 = (0, 0)$ is feasible and optimal for cost vector $c' = (1, -1)$. However, for the point $\bar{z} = \frac{1}{2} z_1 + \frac{1}{2} z_2 = (0, 1)$ there exists no cost vector in $\{1\} \times [-1, 1]$ for which it is optimal. The reason is that for $c_2 > 0$, the point $(0, 2)$ always has a strictly better objective function value, while for $c_2 < 0$, the point $(0, 0)$ always has a strictly better objective function value. For $c_2 = 0$, the point $(1, 1)$ has a strictly better objective function value, since $c_1 = 1$. Hence, the projection onto the x-space is not convex.

Appendix D. Examples for strong counterfactual explanations

The following example shows that the feasible region of the SCEP can be open.

Example 30. Consider the present problem

$$\min \{-x_1 : 0 \leq x_1, x_2 \leq 1\},$$

and assume that only the objective parameters are mutable. Consider the favored solution space $D = \{x : 0.5 \leq x_1 \leq 1,\ 0 \leq x_2 \leq 0.5\}$. The only face of the feasible region which lies entirely in D is the extreme point $(1, 0)$. This extreme point is the unique optimal solution for all $(c_1, c_2) \in (0, \infty) \times (-\infty, 0)$ which is an open set. Hence, the feasible region for (c_1, c_2) of the SCEP for this example is open.

In the following we show that the feasible region of the variables c, A, b may be non-convex and disconnected.

Example 31. Consider again the same present problem in Example 26 together with the same favored solution space. In addition, we assume that only the two objective parameters c_1, c_2 can be changed to any value. Then, the set of objective parameters for which every optimal solution is contained in $D(\hat{x})$ is int $(C_3) \cup$ int (C_4), where C_3 and C_4 are defined as in Example 26. Note that in contrast to Example 27, here for the strong CEs we have to choose the interior of the sets since, *e.g.*, for direction $\binom{1}{1}$ the whole line between points $\binom{0}{1}$ and $\binom{1}{0}$ is optimal, and hence, not every optimal solution is contained in $D(\hat{x})$. Consequently, $\binom{1}{1}$ is not a strong CE. Like in Examples 26 and 27, we can show that the feasible region for the strong CEs is non-convex and disconnected.

Appendix E. Additional results NETLIB experiment

E.1. Instance information NETLIB experiment

Table E.7 contains information on the instance setup. Note that the two columns denoted by "# mutable objective param". and "# mutable constraint param". correspond to the average number of bilinear terms which appear in formulation (2). This value can be different for each NETLIB instance due to the procedure how we select mutable parameters as described in the main text of the paper.

Table E.7
Instance informations for the considered classes of NETLIB instances. We show from left to right the following values: the size of the dimension; the size of the number of constraints; the number of random columns for mutable parameter selection; the number of NETLIB instances which fall into the category; the number of instances for which a feasible relative CE exists in %; the average number of objective parameters which are selected to be mutable; the average number of constraint parameters which are selected to be mutable. All averages are taken over the feasible CE instances.

n	m	# mut. columns	# inst.	feasible (in %)	# mutable objective param.	# mutable constraint param.
Small	Small	1	28	38.00	0.38	4.85
		5	28	54.00	1.67	20.93
		10	28	59.00	4.09	54.67
	Medium	1	7	36.00	0.86	6.09
		5	7	61.00	4.01	32.48
		10	7	64.00	10.78	81.00
Medium	Small	1	4	84.00	0.75	5.88
		5	4	100.00	2.59	22.58
		10	4	100.00	6.75	54.12
	Medium	1	22	35.00	0.47	20.00
		5	22	51.00	2.41	26.66
		10	22	58.00	6.35	42.81
	Large	1	8	31.00	0.56	3.51
		5	8	55.00	2.07	11.92
		10	8	63.00	5.66	27.24
Large	Small	1	2	25.00	1.00	0.00
		5	2	48.00	1.89	0.42
		10	2	50.00	4.05	1.15
	Medium	1	6	43.00	0.63	2.49
		5	6	49.00	3.34	10.46
		10	6	53.00	8.62	26.16
	Large	1	21	45.00	0.60	5.49
		5	21	58.00	2.50	22.75
		10	21	65.00	6.40	54.14

Table E.8
Solution quality information for the considered classes of NETLIB instances. From left to right we show the following information: the size of the dimension; the size of the number of constraints; the number of random columns for mutable parameter selection; the number of objective parameters which are changed compared to the present problem; the number of constraint parameters which are changed compared to the present problem.

			Bilinear formulation (2) with obj. $\sum_{j=1}^{n} \delta_j \left((c, A), (\hat{c}, \hat{A})\right)$		Linear reformulation (10) with obj. $\sum_{j=1}^{n} x_j \delta_j \left((c, A), (\hat{c}, \hat{A})\right)$	
n	m	# mut. col.	$\|c - \hat{c}\|_0$	$\|A - \hat{A}\|_0$	$\|c - \hat{c}\|_0$	$\|A - \hat{A}\|_0$
Small	Small	1	0.25	2.24	0.18	1.14
		5	0.46	4.39	0.20	1.20
		10	0.42	6.08	0.17	1.34
	Medium	1	0.63	2.30	0.55	1.04
		5	0.84	4.56	0.45	2.03
		10	0.76	5.31	0.34	2.88
Medium	Small	1	0.49	2.13	0.38	1.66
		5	0.50	4.08	0.09	2.17
		10	1.06	10.92	0.00	1.99
	Medium	1	0.33	2.59	0.31	0.92
		5	0.42	3.23	0.31	0.98
		10	0.58	4.31	0.32	0.96
	Large	1	0.49	1.37	0.47	0.75
		5	0.63	2.02	0.48	1.02
		10	0.80	3.60	0.53	1.16
Large	Small	1	1.00	0.00	1.00	0.00
		5	1.26	0.00	1.26	0.00
		10	1.45	0.05	1.20	0.05
	Medium	1	0.44	1.05	0.42	0.93
		5	0.55	2.09	0.48	0.97
		10	0.75	6.15	0.71	1.25
	Large	1	0.42	1.73	0.41	1.64
		5	0.56	1.57	0.47	1.39
		10	0.89	1.78	0.69	1.64

Table E.9
Solution time information for the considered classes of NETLIB instances. We show from left to right the following values: the size of the dimension; the size of the number of constraints; the number of random columns for mutable parameter selection; the amount of instances for which the solution method hit the time limit of 1800 s; the solution time t in seconds (averaged over all instances which are not infeasible); the time t_{inf} in seconds which was needed to detect infeasibility (averaged over all instances which are infeasible); the time t in seconds to solve the original problem (PP); the time t_{setup} to set up the problem formulation for the present problem.

			Bilinear formulation (2) with obj. $\sum_{j=1}^{n} \delta_j \left((c, A), (\hat{c}, \hat{A})\right)$			Linear reformulation (10) with obj. $\sum_{j=1}^{n} x_j \delta_j \left((c, A), (\hat{c}, \hat{A})\right)$			Present problem (PP)	
n	m	# mut. col.	Hit TL (%)	t	t_{inf}	Hit TL(%)	t	t_{inf}	t	t_{setup}
Small	Small	1	0.00	0.19	0.13	0.00	0.14	0.13		
		5	1.00	29.19	0.13	0.00	0.15	0.13	0.13	0.12
		10	2.00	43.97	0.14	0.00	0.17	0.15		
	Medium	1	0.00	0.42	0.31	0.00	0.31	0.30		
		5	1.00	16.09	0.31	0.00	0.33	0.31	0.32	0.30
		10	3.00	53.40	0.32	0.00	0.36	0.33		
Medium	Small	1	0.00	0.79	0.78	0.00	0.53	0.77		
		5	0.00	1.49	nan	0.00	0.54	nan	0.53	0.49
		10	4.00	109.16	nan	0.00	0.58	nan		
	Medium	1	0.00	0.99	0.45	0.00	0.50	0.44		
		5	0.00	11.28	0.44	0.00	0.49	0.44	0.50	0.45
		10	1.00	32.76	0.45	0.00	0.51	0.45		
	Large	1	0.00	1.64	1.03	0.00	1.04	1.03		
		5	1.00	21.10	1.00	0.00	1.05	1.01	1.14	1.06
		10	2.00	51.35	1.01	0.00	1.08	1.04		
Large	Small	1	0.00	11.91	8.40	0.00	11.10	8.27		
		5	0.00	11.43	8.48	0.00	10.50	8.28	8.22	7.98
		10	0.00	12.17	7.27	0.00	11.08	7.19		
	Medium	1	0.00	2.17	1.40	0.00	0.90	1.36		
		5	0.00	3.48	1.39	0.00	0.92	1.35	1.34	1.24
		10	0.00	15.65	1.42	0.00	0.95	1.38		
	Large	1	1.00	94.58	28.78	0.00	4.90	5.36		
		5	3.00	117.39	24.37	0.00	4.42	5.40	17.35	3.18
		10	6.00	202.81	27.19	0.00	4.95	4.90		

E.2. Numerical comparison of bilinear and convex formulation for relative CEs

We show the difference in computational tractability in calculating relative CEs by the bilinear formulation (2) and by the convex reformulation (10) with objective (12) using δ_j as the ℓ_1-norm. We use the same setup as for the experiments in Section 5.2.

Table E.8 contains information about the solution quality of both problems. In the solutions of the linear problem (10), the number of changed parameters (both objective and constraint parameters) is smaller than for problem formulation (2). That is, the linear problem leads to sparser changes in the problem parameters. This is probably due to the convexity of the linear problem, since minimizing the ℓ_1-norm over convex sets leads to sparse solutions.

Table E.9 shows results on the runtime of the methods. The results show that for all instance sizes the linear problems can be solved much faster than the bilinear formulation. The improvement in solution time increases for larger instance sizes. The same holds for detecting infeasibility. Furthermore, the bilinear formulation could not be solved to optimality in all cases within the time limit, while the linear problems never hit the time limit. We also show the time to solve the present problem (PP) and the time to set up the problem formulation for the present problem. The latter metric is motivated by our observation that setting up the constraint matrix (although using the sparse-matrix datatype provided by Numpy) takes a significant amount of the runtime for both, the present problem and the counterfactual problems.

References

Aigner, K.-M., Goerigk, M., Hartisch, M., Liers, F., & Miehlich, A. (2024). A framework for data-driven explainability in mathematical optimization. *vol. 38*, In *Proceedings of the AAAI conference on artificial intelligence* (19), (pp. 20912–20920).

Amaral, P., Júdice, J., & Sherali, H. D. (2008). A reformulation–linearization–convexification algorithm for optimal correction of an inconsistent system of linear constraints. *Computers & Operations Research*, *35*(5), 1494–1509.

Barratt, S., Angeris, G., & Boyd, S. (2021). Automatic repair of convex optimization problems. *Optimization and Engineering*, *22*, 247–259.

Bertsimas, D., Delarue, A., Eger, W., Hanlon, J., & Martin, S. (2020). Bus routing optimization helps boston public schools design better policies. *INFORMS Journal on Applied Analytics*, *50*(1), 37–49.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI)*, *8*(1), 8—13.

Bogetoft, P., Ramírez-Ayerbe, J., & Morales, D. R. (2024). Counterfactual analysis and target setting in benchmarking. *European Journal of Operational Research*, *315*(3), 1083–1095.

Brekelmans, R., den Hertog, D., Roos, C., & Eijgenraam, C. (2012). Safe dike heights at minimal costs: The nonhomogeneous case. *Operations Research*, *60*(6), 1342–1355.

Chan, T. C., Mahmood, R., & Zhu, I. Y. (2025). Inverse optimization: Theory and applications. *Operations Research*, *73*(2), 1046–1074.

Dantzig, G. (1963). *Linear programming and extensions*. University Press.

Dempe, S. (2002). *Foundations of bilevel programming*. Springer.

Dempe, S., Kalashnikov, V., Pérez-Valdés, G. A., & Kalashnykova, N. (2015). Bilevel programming problems. *10*, In *Energy Systems* (pp. 3–978). Berlin: Springer.

Eijgenraam, C., Brekelmans, R., den Hertog, D., & Roos, C. (2017). Optimal strategies for flood prevention. *Management Science*, *63*(5), 1644–1656.

EUR-Lex (2016). *Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. European Union, http://data.europa.eu/eli/reg/2016/679/2016-05-04, (Last Accessed 20 March 2024).

EUR-Lex (2021). *Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, (Last Accessed 20 March 2024).

Forel, A., Parmentier, A., & Vidal, T. (2023). Explainable data-driven optimization: from context to decision and back again. In *International Conference on Machine Learning* (pp. 10170–10187). PMLR.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability*. New York: W. H. Freeman and Company.

Goerigk, M., & Hartisch, M. (2023). A framework for inherently interpretable optimization models. *European Journal of Operational Research*, *310*(3), 1312–1324.

Gorissen, B., den Hertog, D., & Reusken, M. (2022). Hidden convexity in a class of optimization problems with bilinear terms. *Optimization Online*, URL https://optimization-online.org/?p=19067.

Kleinert, T., Labbé, M., Ljubić, I., & Schmidt, M. (2021). A survey on mixed-integer programming techniques in bilevel optimization. *EURO Journal on Computational Optimization*, *9*, Article 100007.

Korikov, A., & Beck, J. C. (2021). Counterfactual explanations via inverse constraint programming. In *27th international conference on principles and practice of constraint programming (CP 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Korikov, A., & Beck, J. C. (2023). Objective-based counterfactual explanations for linear discrete optimization. In *International conference on integration of constraint programming, artificial intelligence, and operations research* (pp. 18–34). Springer.

Korikov, A., Shleyfman, A., & Beck, J. C. (2021). Counterfactual explanations for optimization-based decisions in the context of the GDPR. In *ICAPS 2021 workshop on explainable AI planning*.

Krishnakumari, P. K., Antonissen, J., Theulen, F., Gromicho, J., den Hertog, D., Kaiser, K., & Kant, G. (2024). Optimizing geospatial accessibility to healthcare services in low- and middle-income countries. Working Paper.

Kroon, L., Huisman, D., Abbink, E., Fioole, P.-J., Fischetti, M., Maróti, G., Schrijver, A., Steenbeek, A., & Ybema, R. (2009). The new dutch timetable: The OR revolution. *Interfaces*, *39*(1), 6–17.

Maragno, D., Röber, T. E., & Birbil, I. (2022). Counterfactual explanations using optimization with constraint learning. In *OPT 2022: optimization for machine learning (neurIPS 2022 workshop)*.

Maragno, D., Wiberg, H., Bertsimas, D., Birbil, Ş. İ., den Hertog, D., & Fajemisin, A. O. (2025). Mixed-integer optimization with constraint learning. *Operations Research*, *73*(2), 1011–1028.

Marcotte, P., & Savard, G. (2005). Bilevel programming: A combinatorial perspective. In *Graph theory and combinatorial optimization* (pp. 191–217). Springer.

Meijer, P., & van Veluw, S. (2024). The Feed Calculator App. https://www.feedcalculator.org/, (Accessed 8 April 2024).

Moosaei, H., & Hladík, M. (2021). On the optimal correction of infeasible systems of linear inequalities. *Journal of Optimization Theory and Applications*, *190*(1), 32–55.

NETLIB (2024). Netlib repository. https://www.netlib.org/, (Last Accessed 18 April March).

OSTP (2022). *Blueprint for an AI Bill of Rights: Making automated systems work for the American people*. USA: The White House Office of Science and Technology Policy (OSTP, https://www.whitehouse.gov/ostp/ai-bill-of-rights, (Last Accessed 20 April March).

Peters, K., Silva, S., Gonçalves, R., Kavelj, M., Fleuren, H., den Hertog, D., Ergun, O., & Freeman, M. (2021). The nutritious supply chain: Optimizing humanitarian food assistance. *INFORMS Journal on Optimization*, *3*, 200–226.

Peters, K., Silva, S., Wolter, T., Anjos, L., van Ettekoven, N., Combette, É., Melchiori, A., Fleuren, H., den Hertog, D., & Ergun, Ö. (2022). UN world food programme: Toward zero hunger with analytics. *INFORMS Journal on Applied Analytics*, *52*, 8–26.

Rada, M., Hladík, M., & Garajová, E. (2019). Testing weak optimality of a given solution in interval linear programming revisited: NP-hardness proof, algorithm and some polynomially-solvable cases. *Optimization Letters*, *13*, 875–890.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887.

Wang, L. (2013). Branch-and-bound algorithms for the partial inverse mixed integer linear programming problem. *Journal of Global Optimization*, [ISSN: 1573-2916] *55*(3), 491–506.

Wiesemann, W., Tsoukalas, A., Kleniati, P.-M., & Rustem, B. (2013). Pessimistic bilevel optimization. *SIAM Journal on Optimization*, *23*(1), 353–380.