Linear Model Extraction via Factual and Counterfactual Queries

Daan Otto
d.otto@uva.nl

Jannis Kurtz
j.kurtz@uva.nl

Dick den Hertog
d.denhertog@uva.nl

Ilker Birbil
s.i.birbil@uva.nl
Amsterdam Business School

University of Amsterdam
Amsterdam, 1001 NL, The Netherlands

Abstract

In model extraction attacks, the goal is to reveal the parameters of a black-box machine learning model by querying the model for a selected set of data points. Due to an increasing demand for explanations, this may involve counterfactual queries besides the typically considered factual queries. In this work, we consider linear models and three types of queries: factual, counterfactual, and robust counterfactual. First, for an arbitrary set of queries, we derive novel mathematical formulations for the classification regions for which the decision of the unknown model is known, without recovering any of the model parameters. Second, we derive bounds on the number of queries needed to extract the model's parameters for (robust) counterfactual queries under arbitrary norm-based distances. We show that the full model can be recovered using just a single counterfactual query when differentiable distance measures are employed. In contrast, when using polyhedral distances for instance, the number of required queries grows linearly with the dimension of the data space. For robust counterfactuals, the latter number of queries doubles. Consequently, the applied distance function and robustness of counterfactuals have a significant impact on the model's security.

1 Introduction

As machine learning models become progressively more prevalent in research and real-world applications, there is an increasing attention to their impact on privacy, security, and explainability. Recent work outlines attack techniques that threaten model security and privacy, such as model extraction [Rigaki and Garcia, 2023]. Model extraction aims to reconstruct a target black-box model by querying specific data points and using their outcomes to find the original model's parameters or train a surrogate model that replicates the original model's behavior. Model extraction attacks can jeopardize the model integrity and the intellectual property of the model owner. In combination with model inversion attacks [Fredrikson et al., 2015] or attacks that reconstruct training data [Boenisch et al., 2023, Ferry et al., 2024], this poses privacy risks, which are especially relevant when models are trained on sensitive data, *e.g.*, in medical or financial domains.

Besides, as machine learning models become more complex, their inner workings become less comprehensible to human users. This decrease in explainability leads to the continuously growing field of Explainable Artificial Intelligence (XAI), providing tools for explanations for a large variety of machine learning models, such as counterfactual explanations [Wachter et al., 2017]. A counterfactual for a given factual instance is a (small) perturbation of the instance itself such that the decision of

the model flips, answering the question: *"In what situation would the outcome be B instead of A?"*. However, as counterfactual explanations can enhance transparency and trustworthiness, significant security risks are introduced, as the explanations can expose sensitive information about data and the underlying black-box model [Shokri et al., 2021, Nguyen et al., 2024, Milli et al., 2019, Oksuz et al., 2024]. Attackers may exploit counterfactuals to extract the true model parameters [Khouna et al., 2025].

Model extraction methods using factual or counterfactual queries were already studied in several works. Lowd and Meek [2005] demonstrated that the parameters of linear classifiers on continuous data can be extracted within an $(1+\epsilon)$ factor using a polynomial number of factual queries. Tramèr et al. [2016] extend on this and present model extraction attacks for other model classes, including logistic regression, neural networks, and decision trees. The authors consider the machine learning-as-a-service setting, where partial feature vectors can also be queried to obtain confidence values for the model predictions. Besides, Reith et al. [2019] focus on model extraction of Support Vector Regression models using equation-solving attacks. By querying data points, they find a system of equations that find the model parameters. Another approach that has been studied tries to minimize the query cost by framing model extraction as an active learning problem [Chandrasekaran et al., 2020, Pal et al., 2020].

Other works additionally consider counterfactual queries. In Berning et al. [2024] and Goethals et al. [2023], the authors describe the privacy issues when the counterfactual mechanism uses an actual existing data point, *e.g.*, another person's data. These works present how a trade-off between privacy and the quality of counterfactual explanations can be made using an anonymization method. A different approach is presented by Aïvodji et al. [2020] and consists of creating a dataset of factual instances, then querying a set of counterfactual queries to obtain a balanced dataset to train a surrogate model. In their work, Wang et al. [2022] propose a two-step approach that uses counterfactuals of counterfactuals to obtain the parameters of a linear classifier. Here, they assume the counterfactuals are not lying on the decision boundary. They show the effectiveness of their model experimentally. However, no theoretical bounds on the number of queries needed to extract the original model are presented. Dissanayake and Dutta [2024] use the ℓ_2-norm for the *minimal edit* counterfactual to obtain polytope approximations of classifiers whose decision boundary is convex and has a continuous second derivative. The authors also discuss the reconstruction of neural networks with ReLU activation functions using counterfactual queries. Lastly, Khouna et al. [2025] use counterfactual queries to reconstruct decision trees exactly and provide guarantees on the complexity of the number of queries required for model extraction. However, for many model classes, theoretical bounds on the required number of queries for complete model extraction are still not researched.

In this work, we focus on linear models and consider three types of queries: (i) factuals, (ii) exact counterfactuals, and (iii) exact robust counterfactuals. We study the research questions (i) how much arbitrary queries reveal about the classification regions, and (ii) how many queries are needed to recover the model parameters. By applying techniques from robust optimization, we first derive novel mathematical formulations for the classification regions for the situation where an arbitrary set of query results is given. This extends the current literature to the situation where hand-crafted queries cannot be performed. Second, we show that with a small number of targeted queries the exact model parameters can be recovered. The derived bounds extend the current literature by considering more general setups of counterfactuals, involving arbitrary norm-based distance functions, and robust counterfactuals. Our developed theory shows the effect of the distance function and the robustness of the counterfactuals on the number of required queries for model extraction.

While the performance of linear models in machine learning is restricted due to their limited complexity, they are widely used because of their interpretability. In highly regulated fields like banking, there are often restrictions on using nonlinear predictors. Regulatory frameworks such as the BCBS [Basel Committee on Banking Supervision, 2026], the GDPR's rules on automated decision-making [European Union, 2016], and the SR 11-7 Guidance on Model Risk Management [Federal Reserve System, 2011] emphasize explainability, documentation, and transparency, which

favors the adoption of inherently interpretable models like linear regression. Besides, the simple structure of linear models allows us to derive precise mathematical formulations for the subsets of the classification regions and exact bounds on the number of (robust) counterfactual queries needed to extract the model parameters exactly. This analysis provides a foundation that may be extended to more complex classifiers in future work.

Our contributions consist of the following:

1. We derive novel and computationally tractable characterizations for the data points for which we can detect the classification without querying the model again, when provided an arbitrary set of (i) factuals, (ii) exact counterfactuals, or (iii) exact robust counterfactuals.

2. We extend on the current literature by providing upper bounds on the number of (robust) counterfactual queries needed to fully extract the linear classifier's parameters for general norm-based distance functions. Our results show that the choice of the distance measure for (robust) counterfactuals has a significant impact on the number of queries needed for model extraction.

2 Preliminaries

In this work, we denote vectors in boldface. Subscripts are used to index vector components, while superscripts are used to index different vectors. We denote the standard basis vectors of $\mathbb{R}^p$ with $\boldsymbol{e}^1, \ldots, \boldsymbol{e}^p$. We use superscripts in parentheses, $\boldsymbol{x}^{(i)}$ to denote data points in a collection $\{\boldsymbol{x}^{(i)} \mid i \in I\}$.

Consider a trained classifier $h_{\boldsymbol{a},b} : \mathcal{X} \to \{-1, 1\}$ which maps each data point $\boldsymbol{x}$ in the data space $\mathcal{X} \subseteq \mathbb{R}^p$ either to class -1 (*'No'*) or 1 (*'Yes'*). Concretely, we consider linear classifiers given by the hyperplane of the form $\boldsymbol{a}^\top \boldsymbol{x} - b = 0$ such that

$$h_{\boldsymbol{a},b}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{a}^\top \boldsymbol{x} - b \geq 0; \\ -1, & \text{if } \boldsymbol{a}^\top \boldsymbol{x} - b < 0, \end{cases}$$

where $\boldsymbol{a} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}, b \in \mathbb{R}$. We denote $(\boldsymbol{a}, b)$ as the parameter vector of the linear classifier h. Note that classification models involving non-linear transformations of a linear model of the form $f(\boldsymbol{a}^\top \boldsymbol{x}) \geq \alpha$ fall in our framework if f is monotonic and invertible, since in this case we can equivalently reformulate $\boldsymbol{a}^\top \boldsymbol{x} \geq f^{-1}(\alpha)$ and set $b := f^{-1}(\alpha)$. Hence, all results in this work can be applied to logistic regression where f is the sigmoid function.

We make the following general assumptions:

(A1) We assume that $\mathcal{X}$ is the full real data space, *i.e.*, $\mathcal{X} = \mathbb{R}^p$.

(A2) We assume a non-zero $\boldsymbol{a}$, *i.e.*, there exists an $i \in \{1, \ldots, p\}$ with $a_i \neq 0$. Hence, both *'Yes'* and *'No'* points exist.

Evidently, two hyperplanes given by $(\boldsymbol{a}, b)$ and $(\hat{\boldsymbol{a}}, \hat{b})$ are equivalent iff there exists a non-zero scalar λ such that $(\boldsymbol{a}, b) = \lambda(\hat{\boldsymbol{a}}, \hat{b})$. Hence, hyperplanes are invariant under scaling.

Definition 1 (Equivalent Hyperplane)**.** *Given a hyperplane with parameters $(\boldsymbol{a}, b)$ such that $\boldsymbol{a}^\top \boldsymbol{x} - b = 0$, an equivalent hyperplane is given by $(\hat{\boldsymbol{a}}, \hat{b})$ such that $\hat{\boldsymbol{a}}^\top \boldsymbol{x} - \hat{b} = 0$ if and only if $\boldsymbol{a}^\top \boldsymbol{x} - b = 0$.*

Next, we define the different query mechanisms we consider in this work, (i) factual, (ii) counterfactual, and (iii) robust counterfactual, which are also depicted in Fig. 1.

Definition 2 (Factual Query)**.** *A factual query $q_F : \mathcal{X} \to \{0, 1\}$ maps a data point $\boldsymbol{x} \in X$ to the label of the linear classifier, i.e., $q_F(\boldsymbol{x}) := h_{\boldsymbol{a},b}(\boldsymbol{x})$.*

A mechanism often used in XAI is the counterfactual query. This mechanism outputs a *minimal edit* to a data point to get a desired output from the original model.

Definition 3 (Counterfactual Query)**.** *Given an arbitrary norm $\|\cdot\|_{N_1}$ (norm-1), a counterfactual (CF) query $q_{CF} : \mathcal{X} \to \mathcal{X}$ maps a data point $\boldsymbol{x} \in \mathcal{X}$ to an optimal solution $\boldsymbol{x}^*_{CF}$ of the problem*

$$\begin{aligned} \min_{\boldsymbol{x}_{CF}} \ & \|\boldsymbol{x}_{CF} - \boldsymbol{x}\|_{N_1} \\ s.t. \quad & h_{\boldsymbol{a},b}(\boldsymbol{x}) \neq h_{\boldsymbol{a},b}(\boldsymbol{x}_{CF}), \\ & \boldsymbol{x}_{CF} \in \mathcal{X}. \end{aligned} \tag{1}$$

Note that with slight abuse of notation, we use the minimum operator instead of the infimum operator in the latter problem. Since the region for points which are classified as *'No'* is open, it may happen that the latter optimization problem has no optimal solution. However, in practical settings, a counterfactual is usually calculated over the closure of the region, both for points classified as *'Yes'* or *'No'*. Therefore, when defining the classification regions or calculating (robust) counterfactuals we will consider the closure of the classification regions.

A drawback of counterfactual queries is the lack of robustness; a counterfactual lies on a decision boundary, meaning it is highly sensitive to slight changes in the data point. To combat this problem, we also consider robust counterfactuals which were proposed in the literature, *e.g.*, see Maragno et al. [2024].

Definition 4 (Robust Counterfactual Query)**.** *For a given robustness set $\mathcal{S}$ and an arbitrary norm $\|\cdot\|_{N_1}$, a robust counterfactual (RCF) query $q_{RCF} : \mathcal{X} \to \mathcal{X}$ maps a data point $\boldsymbol{x} \in \mathcal{X}$ to an optimal solution $\boldsymbol{x}^*_{RCF}$ of the problem*

$$\begin{aligned} \min_{\boldsymbol{x}_{RCF}} \ & \|\boldsymbol{x}_{RCF} - \boldsymbol{x}\|_{N_1} \\ s.t. \quad & h_{\boldsymbol{a},b}(\boldsymbol{x}) \neq h_{\boldsymbol{a},b}(\boldsymbol{x}_{RCF} + \boldsymbol{s}) \quad \forall \boldsymbol{s} \in \mathcal{S}, \\ & \boldsymbol{x}_{RCF} \in \mathcal{X}. \end{aligned} \tag{2}$$

The definition of a robust counterfactual ensures that for each perturbation of the point by a point in the robustness set $\mathcal{S}$, the perturbed point remains a counterfactual. Note that a common class of robustness sets is the class of norm-balls of a given radius, where the norm does not have to coincide with the norm used in the objective function of problem (2). To prevent confusion, we use $\|\cdot\|_{N_2}$ (norm-2) to define the robustness set, *i.e.*, $\mathcal{S} := \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$ for $\rho > 0$. Geometrically, a robust counterfactual is the closest point to the factual instance, such that the whole norm-2-ball around the point lies on the other side of the decision boundary; see Figure 1.

In the following, we distinguish between norms that are differentiable at any point $\boldsymbol{x} \in \mathcal{X} \setminus \{0\}$ and norms that do not have this property. We will respectively refer to these norms as differentiable norms and non-differentiable norms. Examples of differentiable norms are ℓ_p-norms with $1 < p < \infty$, while ℓ_p-norms with $p \in \{1, \infty\}$ are examples for non-differentiable norms.

Table 1 presents a summary of our key results on model extraction using the various query mechanisms introduced above. From this table, we can conclude that using a non-differentiable norm-1 preserves privacy more than a differentiable norm. Moreover, when using robust counterfactuals, we see that more queries are needed to extract the model's parameters than for regular counterfactuals.

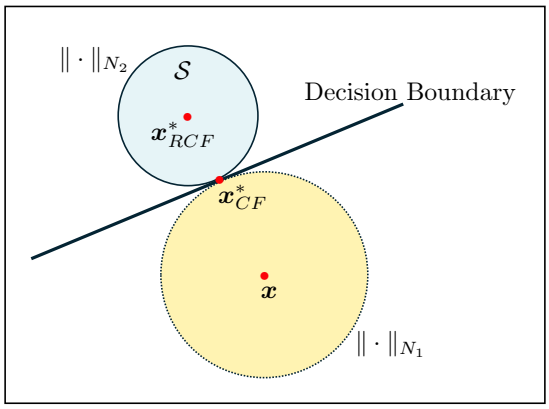


Figure 1: Illustration of the definitions.

Query Type	Norm-1 differentiable	Reconstruction	#Queries	Result
Factual	-	ϵ-approx.	$\mathcal{O}(\log(\epsilon^{-1}) + \text{size}(c))$	Lowd and Meek [2005]
CF	Yes	Exact	1	Theorem 8
CF	No	Exact	$p+1$	Theorem 14
RCF	Yes	Exact	1 RCF & 1 Factual	Theorem 17
RCF	No	Exact	$p+1$ RCFs & $p+1$ Factuals	Corollary 21

Table 1: Information of hyperplane extraction by different query types. Lowd and Meek [2005] assume the model parameters have magnitude either 0 or in $[2^{-c}, 2^c]$.

3 Factual Queries

In Lowd and Meek [2005], it was shown that by smartly performing up to $\log(\epsilon^{-1})$ factual queries for the classifier $h_{\boldsymbol{a},b}$, the hyperplane parameters $\boldsymbol{a}, b$ can be recovered up to relative accuracy of $\epsilon > 0$. However, the latter result relies on the situation that we are able to query a potentially large number of points. On the other hand, if only an arbitrary number of query results is given and we do not have the possibility to perform the required amount of queries, the question remains whether we can already know the query outcome of certain points without querying the model. For example, if we have a set of points given for which we know the model returns the classification 1, then we know that every point in the convex hull of the given points must be classified as 1 as well. We next discuss how much information is extracted regarding the *'Yes'* and *'No'* regions by an arbitrary set of factual queried data points and show novel mathematical formulations for these regions which extend beyond the convex hull.

Suppose we have a set of data points $\{\boldsymbol{x}^{(i)}, i \in I\}$ where I is the index set. Moreover, assume that the output of the factual query for $\boldsymbol{x}^{(i)}$ is *'No'* for $i \in I_0$, and is *'Yes'* for $i \in I_1$, where $I_0 \cup I_1 = I$. We then get the following convex region (polyhedron) as possible values for the unknown parameters $\boldsymbol{a}, b$:

$$\mathcal{U}^F_{\boldsymbol{a},b} = \{(\boldsymbol{a}, b) \mid \boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b \leq 0 \; \forall i \in I_0, \quad \boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b \geq 0 \; \forall i \in I_1\}. \tag{3}$$

We now consider the question: For a given data point $\bar{\boldsymbol{x}} \in \mathcal{X}$, can we already know –based on the information given by the factual queries– whether the classifier yields a *'Yes'* or *'No'*? Clearly, each point in the convex hull of $\boldsymbol{x}^{(i)}, \forall i \in I_0$, will yield a *'No'*, and each point in the convex hull of $\boldsymbol{x}^{(i)}, \forall i \in I_1$, will yield a *'Yes'*.

However, we can show that there are many more data points for which we can detect their classification without querying the model again. We refer to all data points for which we can detect the classification will be *'No'* without querying the model again as the *'No'* region ($\mathcal{X}_{\text{'No'}}$). Similarly, the *'Yes'* region ($\mathcal{X}_{\text{'Yes'}}$) refers to all data points for which we can detect that the classification will be *'Yes'* without querying the model again. We can test whether we will obtain a *'No'* for $\bar{\boldsymbol{x}}$ by solving the following linear optimization problem:

$$\max_{\boldsymbol{a},b}\{\boldsymbol{a}^\top \bar{\boldsymbol{x}} - b \ : \ (\boldsymbol{a},b) \in \mathcal{U}^F_{\boldsymbol{a},b}\}.$$

If the optimal value of this problem is at most 0, this means that for all possible $(\boldsymbol{a},b) \in \mathcal{U}^F_{\boldsymbol{a},b}$, the original classifier will output a *'No'* for $\bar{\boldsymbol{x}}$. This results in the following convex set for the *'No'* region:

$$\mathcal{X}_{\text{'No'}} := \{\boldsymbol{x} \ \mid \ \boldsymbol{a}^\top \boldsymbol{x} - b \leq 0 \quad \forall \boldsymbol{a}, b \in \mathcal{U}^F_{\boldsymbol{a},b}\} = \{\boldsymbol{x} \ \mid \ \max_{\boldsymbol{a},b \in \mathcal{U}^F_{\boldsymbol{a},b}} \boldsymbol{a}^\top \boldsymbol{x} - b \leq 0\}. \tag{4}$$

On the other hand, if there exists an $(\boldsymbol{a},b) \in \mathcal{U}^F_{\boldsymbol{a},b}$ such that $\boldsymbol{a}^\top \bar{\boldsymbol{x}} - b > 0$ we cannot know whether $\bar{\boldsymbol{x}}$ will be classified as *'No'* without factual querying $\bar{\boldsymbol{x}}$. Similarly, if the optimal value of the following linear optimization problem

$$\min_{\boldsymbol{a},b}\{\boldsymbol{a}^\top \bar{\boldsymbol{x}} - b \ : \ (\boldsymbol{a},b) \in \mathcal{U}^F_{\boldsymbol{a},b}\}$$

is larger than 0, the original classifier will output *'Yes'* for $\bar{\boldsymbol{x}}$. Moreover, note that the set of points, for which we know for sure that we will obtain a *'Yes'* from the original classifier, is the following convex region:

$$\mathcal{X}_{\text{'Yes'}} := \{\boldsymbol{x} \ \mid \ \boldsymbol{a}^\top \boldsymbol{x} - b \geq 0 \quad \forall \boldsymbol{a}, b \in \mathcal{U}^F_{\boldsymbol{a},b}\} = \{\boldsymbol{x} \ \mid \ \min_{\boldsymbol{a},b \in \mathcal{U}^F_{\boldsymbol{a},b}} \boldsymbol{a}^\top \boldsymbol{x} - b \geq 0\}. \tag{5}$$

By dualizing the optimization problems described in (4) and (5), we get different equivalent formulations of the *'Yes'* and *'No'* regions. In the following theorem, we derive such a mathematical formulation for the regions for which we can detect the classification without additional queries.

Theorem 5. *Given data points $\boldsymbol{x}^{(i)}$, $i \in I$ such that each $\boldsymbol{x}^{(i)}$ is classified as 'No' for $i \in I_0$ and as 'Yes' for $i \in I_1$, where $I_0 \cup I_1 = I$. Then, the 'No' and 'Yes' regions are given by*

$$\mathcal{X}_{\text{'No'}} = \left\{ \boldsymbol{x} \mid \ \exists \boldsymbol{u} \ : \ \sum_{i \in I_0} u_i - \sum_{i \in I_1} u_i = 1, \ \sum_{i \in I_0} \boldsymbol{x}^{(i)} u_i - \sum_{i \in I_1} \boldsymbol{x}^{(i)} u_i = \boldsymbol{x}, \ \boldsymbol{u} \geq \boldsymbol{0} \right\},$$

and

$$\mathcal{X}_{\text{'Yes'}} = \left\{ \boldsymbol{x} \mid \ \exists \boldsymbol{u} \ : \ \sum_{i \in I_1} u_i - \sum_{i \in I_0} u_i = 1, \ \sum_{i \in I_1} \boldsymbol{x}^{(i)} u_i - \sum_{i \in I_0} \boldsymbol{x}^{(i)} u_i = \boldsymbol{x}, \ \boldsymbol{u} \geq \boldsymbol{0} \right\},$$

respectively.

The results of Theorem 5 show that finding out whether a data point $\bar{\boldsymbol{x}}$ is in $\mathcal{X}_{\text{'No'}}$ or $\mathcal{X}_{\text{'Yes'}}$ is computationally tractable since this can be done by optimizing a trivial objective function over the feasible set described in Theorem 5 resulting in a linear optimization problem which can be solved by state-of-the-art optimization solvers efficiently. Moreover, we notice that if we set $u_i = 0$, $\forall i \in I_1$, then the *'No'* region boils down to the convex hull of the points $\boldsymbol{x}^{(i)}$, $i \in I_0$. An example of *'Yes'* and *'No'* regions is depicted in Fig. 2.

4 Counterfactual Queries

In this section, we first discuss how much information is extracted about the *'Yes'* and *'No'* regions by a set of factual and counterfactual data points. Afterwards, we examine how many counterfactual queries are needed to retrieve the original hyperplane exactly.

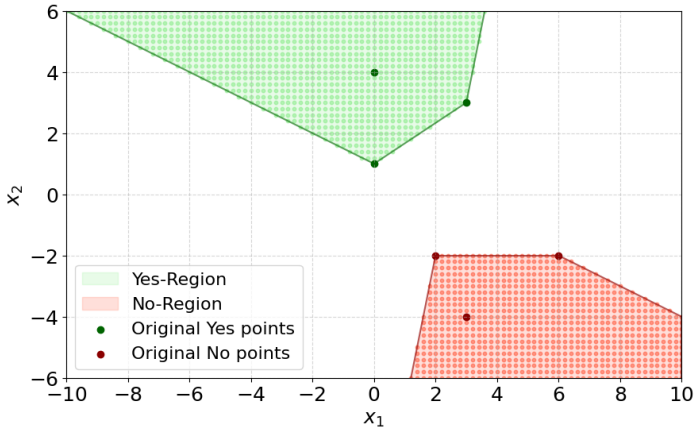


Figure 2: Example of *'Yes'* and *'No'* regions for a set of given factuals.

4.1 Classification Regions

Suppose that, besides factual queries, we also get (*minimal edit*) counterfactuals for some points of our dataset. Concretely, for points $\boldsymbol{x}^{(j)}$ from an index subset $j \in J_0 \cup J_1 = J \subseteq I$ we get a counterfactual $\boldsymbol{x}_{CF}^{(j)} := q_{CF}(\boldsymbol{x}^{(j)})$. Suppose the *minimal edit* is with respect to a certain norm $\|\cdot\|_{N_1}$, and let $\rho_j = \|\boldsymbol{x}^{(j)} - \boldsymbol{x}_{CF}^{(j)}\|_{N_1}$. Then, we know that each point $\boldsymbol{z}$ in the ball with center $\boldsymbol{x}^{(j)}$ and radius ρ_j will be classified similarly to $\boldsymbol{x}^{(j)}$, *i.e.*, this ball lies on the same side of the hyperplane as $\boldsymbol{x}^{(j)}$. Hence, we know

$$\boldsymbol{a}^\top \boldsymbol{z} - b \leq 0 \qquad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}^{(j)}\|_{N_1} \leq \rho_j \qquad \forall j \in J_0, \tag{6}$$

$$\boldsymbol{a}^\top \boldsymbol{z} - b \geq 0 \qquad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}^{(j)}\|_{N_1} \leq \rho_j \qquad \forall j \in J_1. \tag{7}$$

Moreover, we know that a counterfactual has a classification that is opposite to its corresponding factual. Concretely, this means

$$\boldsymbol{a}^\top \boldsymbol{x}_{CF}^{(j)} - b \geq 0 \qquad \forall j \in J_0, \tag{8}$$

$$\boldsymbol{a}^\top \boldsymbol{x}_{CF}^{(j)} - b \leq 0 \qquad \forall j \in J_1. \tag{9}$$

Since constraints (6) and (7) also hold for $\boldsymbol{z} = \boldsymbol{x}_{CF}^{(j)}$, the inequalities in (8) and (9) can be replaced by equalities. With this extra information, we can characterize the uncertainty set of $\boldsymbol{a}, b$ as $\mathcal{U}_{\boldsymbol{a},b}^{CF}$ given by the following set of constraints:

$$\begin{aligned}
\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b &\leq 0 && && \forall i \in I_0,\\
\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b &\geq 0 && && \forall i \in I_1,\\
\boldsymbol{a}^\top \boldsymbol{z} - b &\leq 0 && \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}^{(j)}\|_{N_1} \leq \rho_j && \forall j \in J_0,\\
\boldsymbol{a}^\top \boldsymbol{z} - b &\geq 0 && \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}^{(j)}\|_{N_1} \leq \rho_j && \forall j \in J_1,\\
\boldsymbol{a}^\top \boldsymbol{x}_{CF}^{(j)} - b &= 0 && && \forall j \in J.
\end{aligned}$$

Similar to the case with the factuals, we can write the *'Yes'* and *'No'* region as

$$\mathcal{X}_{\text{'Yes'}} = \{\boldsymbol{x} \mid \min_{\boldsymbol{a},b \in \mathcal{U}_{\boldsymbol{a},b}^{CF}} \boldsymbol{a}^\top \boldsymbol{x} - b \geq 0\} \qquad \text{and} \qquad \mathcal{X}_{\text{'No'}} = \{\boldsymbol{x} \mid \max_{\boldsymbol{a},b \in \mathcal{U}_{\boldsymbol{a},b}^{CF}} \boldsymbol{a}^\top \boldsymbol{x} - b \leq 0\},$$

respectively.

Using this new formulation of the uncertainty set $\mathcal{U}_{\boldsymbol{a},b}^{CF}$, we need perspective functions to dualize the inner optimization problem to obtain a dual characterization for the *'Yes'* and *'No'* regions.

Theorem 6. *Consider a dataset of points $\boldsymbol{x}^{(i)}$ for $i \in I$ such that $q_F(\boldsymbol{x}^{(i)}) = -1$ for all $i \in I_0$ and $q_F(\boldsymbol{x}^{(i)}) = 1$ for all $i \in I_1 = I \setminus I_0$. Moreover, consider the points $\boldsymbol{x}_{CF}^{(j)}$ for $j \in J \subseteq I$ such that $\boldsymbol{x}_{CF}^{(j)} = q_{CF}(\boldsymbol{x}^{(j)})$. Let $J_0 \subseteq I_0, J_1 \subseteq I_1$, and $J = J_0 \cup J_1$. Then, the 'No' and 'Yes' regions are characterized by the conic quadratic sets*

$$\mathcal{X}_{\text{`No'}} = \left\{ \begin{array}{ll} \boldsymbol{x} \mid \exists \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y}: & -\sum_{i \in I_0} t_i + \sum_{i \in I_1} t_i - \sum_{j \in J} y_b - \sum_{j \in J} v_j = -1, \\ & \sum_{i \in I_0} t_i \boldsymbol{x}^{(i)} - \sum_{i \in I_1} t_i \boldsymbol{x}^{(i)} + \sum_{j \in J} \boldsymbol{y}_a^{(j)} + \sum_{j \in J} v_j \boldsymbol{x}_{CF}^{(j)} = \boldsymbol{x}, \\ & u_j(\|\boldsymbol{y}_a^{(j)}/u_j - \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 \quad \forall j \in J_0, \\ & u_j(y_b^{(j)}/u_j + 1) = 0 \quad \forall j \in J_0, \\ & u_j(\|\boldsymbol{y}_a^{(j)}/u_j + \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 \quad \forall j \in J_1, \\ & u_j(y_b^{(j)}/u_j - 1) = 0 \quad \forall j \in J_1, \\ & \boldsymbol{y}^{(j)} = (\boldsymbol{y}_a^{(j)}, y_b^{(j)}) \in \mathbb{R}^{p+1} \quad \forall j \in J \\ & \boldsymbol{t} \in \mathbb{R}_{\geq 0}^{|I|},\ \boldsymbol{u} \in \mathbb{R}_{\geq 0}^{|J|},\ \boldsymbol{v} \in \mathbb{R}^{|J|}. \end{array} \right\}$$

and

$$\mathcal{X}_{\text{`Yes'}} = \left\{ \begin{array}{ll} \boldsymbol{x} \mid \exists \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y}: & -\sum_{i \in I_0} t_i + \sum_{i \in I_1} t_i - \sum_{j \in J} y_b - \sum_{j \in J} v_j = 1, \\ & \sum_{i \in I_0} t_i \boldsymbol{x}^{(i)} - \sum_{i \in I_1} t_i \boldsymbol{x}^{(i)} + \sum_{j \in J} \boldsymbol{y}_a^{(j)} + \sum_{j \in J} v_j \boldsymbol{x}_{CF}^{(j)} = -\boldsymbol{x}, \\ & u_j(\|\boldsymbol{y}_a^{(j)}/u_j - \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 \quad \forall j \in J_0, \\ & u_j(y_b^{(j)}/u_j + 1) = 0 \quad \forall j \in J_0, \\ & u_j(\|\boldsymbol{y}_a^{(j)}/u_j + \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 \quad \forall j \in J_1, \\ & u_j(y_b^{(j)}/u_j - 1) = 0 \quad \forall j \in J_1, \\ & \boldsymbol{y}^{(j)} = (\boldsymbol{y}_a^{(j)}, y_b^{(j)}) \in \mathbb{R}^{p+1} \quad \forall j \in J \\ & \boldsymbol{t} \in \mathbb{R}_{\geq 0}^{|I|},\ \boldsymbol{u} \in \mathbb{R}_{\geq 0}^{|J|},\ \boldsymbol{v} \in \mathbb{R}^{|J|}. \end{array} \right\},$$

respectively.

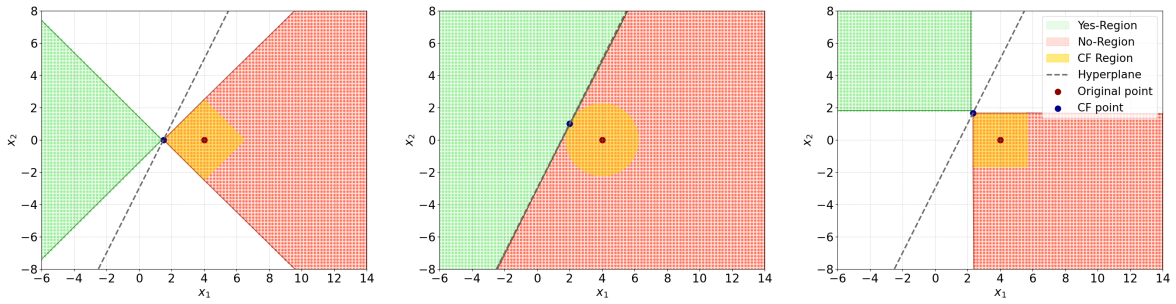
Theorem 6 shows that finding out whether a data point $\bar{\boldsymbol{x}}$ is in $\mathcal{X}_{\text{`No'}}$ or $\mathcal{X}_{\text{`Yes'}}$ is computationally tractable, because of its conic quadratic formulation. To find out whether a data point $\bar{\boldsymbol{x}}$ is in $\mathcal{X}_{\text{`No'}}$ or $\mathcal{X}_{\text{`Yes'}}$, we can optimize a trivial objective function over the set of constraints as described in Theorem 6. The resulting conic quadratic optimization problem can be efficiently solved using state-of-the-art solvers. In Figure 3, we show a two-dimensional visualization of the classification regions when there is one data point and a corresponding counterfactual, *i.e.*, $|I| = |J| = 1$. We consider three cases, when the counterfactual mechanism uses the (i) ℓ_1-norm, (ii) ℓ_2-norm, and (iii) ℓ_∞-norm. We see that for the non-differentiable norms (ℓ_1, ℓ_∞) there are still areas for which we cannot conclude the classification. For the differentiable ℓ_2-norm, however, it seems only one factual and corresponding counterfactual extracts the whole model. In the next section, we verify this hypothesis.

4.2 Extracting the Hyperplane

We will show in this section that the number of counterfactual queries needed to recover the linear classifier $h_{\boldsymbol{a},b}$ depends on the norm $\|\cdot\|_{N_1}$ used in problem (1). We first use classical optimality conditions to derive general conditions on the linear hyperplane parameters $\boldsymbol{a}$ and b.

In the following we denote by $\partial f(\boldsymbol{x}_0)$ the subdifferential set of a function $f : \mathbb{R}^p \to \mathbb{R}$ at point $\boldsymbol{x}_0$, *i.e.*,

$$\partial f(\boldsymbol{x}_0) := \{\boldsymbol{v} \in \mathbb{R}^p : f(\boldsymbol{x}_0) - f(\boldsymbol{x}) \geq \boldsymbol{v}^\top(\boldsymbol{x}_0 - \boldsymbol{x})\ \forall\ \boldsymbol{x} \in \mathcal{X}\}.$$



(a) $N_1 = \ell_1$. (b) $N_1 = \ell_2$. (c) $N_1 = \ell_\infty$.

Figure 3: Example of the classification regions given one data point classified as *'No'* and corresponding counterfactual for different choices of norm-1.

Lemma 7. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $q_{CF}(\boldsymbol{x}_F) = \boldsymbol{x}^*_{CF}$ with $\boldsymbol{x}^*_{CF} \neq \boldsymbol{x}_F$ be its corresponding optimal counterfactual under an arbitrary norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$. Then, there exists a scalar $\lambda^* \in \mathbb{R} \setminus \{0\}$ such that*

$$\begin{aligned} &\boldsymbol{a}^\top \boldsymbol{x}^*_{CF} = b, \\ &\lambda^* \boldsymbol{a} \in \partial f(\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F). \end{aligned}$$

The latter lemma indicates that if f is differentiable, then ∂f is a singleton, which provides us the direction of $\boldsymbol{a}$, while for non-differentiable norms ∂f may be an infinite set, concealing the true direction of $\boldsymbol{a}$. We analyze both cases in the following two subsections.

4.2.1 Differentiable Norms

For differentiable norms, the subdifferential set in Lemma 7 contains only the gradient of f. Hence, in this case $\boldsymbol{a}$ can directly be extracted, which shows that we only need one counterfactual query to recover the hyperplane of the classifier.

Theorem 8. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $q_{CF}(\boldsymbol{x}_F) = \boldsymbol{x}^*_{CF}$ with $\boldsymbol{x}^*_{CF} \neq \boldsymbol{x}_F$ be its corresponding optimal counterfactual under an arbitrary differentiable norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$. This one counterfactual query is enough to extract the original classifier's parameters since for $\hat{\boldsymbol{a}} = \nabla f(\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F)$ and $\hat{b} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}^*_{CF}$ it holds that the hyperplane given by $\hat{\boldsymbol{a}}, \hat{b}$ is equivalent to the original hyperplane with parameters $\boldsymbol{a}, b$.*

Note that Theorem 8 finds a hyperplane equivalent to the one used by the original classifier, however to obtain an equivalent classifier, we still need to know which side of the hyperplane is classified as 1 and which one as -1. This can be done by one factual query of a point lying outside of the hyperplane. Note that the assumption $\boldsymbol{x}_F \neq \boldsymbol{x}^*_{CF}$ in Theorem 8 is not too restrictive since it is only violated if $\boldsymbol{x}_F$ lies on the hyperplane. However, for a randomly drawn point $\boldsymbol{x}_F \in \mathcal{X}$, this happens with probability zero. In case we have a point $\boldsymbol{x}_F$ which lies on the hyperplane, we may query p linearly independent points around $\boldsymbol{x}_F$, *e.g.*, the points $\boldsymbol{x}_F + \boldsymbol{e}^i$ for $i = 1, \ldots, p$. Then at least one of these points does not lie on the hyperplane, and we can apply Theorem 8 to recover the hyperplane.

4.2.2 Non-differentiable norms

Unfortunately, Lemma 7 implies that, in general, one counterfactual query does not extract the original hyperplane if the norm $\|\cdot\|_{N_1}$ is non-differentiable. This is because the subdifferential for non-differentiable norms is not necessarily a singleton but an infinite set. Figures 3a and 3c show,

in the case of non-differentiable norms, that the uncertainty in $\boldsymbol{a}, b$ leads to regions for which the classification is not known. In this case, the subdifferential set of the norm-function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$ in Lemma 7 can be reformulated as

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{v} \in \mathbb{R}^p : \boldsymbol{v}^\top \boldsymbol{x} = \|\boldsymbol{x}\|_{N_1}, \|\boldsymbol{v}\|_{N_1}^* \leq 1\},$$

where $\|\cdot\|_{N_1}^*$ is the dual norm; see Rockafellar [1997]. This leads to the following result.

Corollary 9. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $q_{CF}(\boldsymbol{x}_F) = \boldsymbol{x}_{CF}^*$ under an arbitrary norm $\|\cdot\|_{N_1}$. Then, $\boldsymbol{x}_{CF}^*$ is an optimal counterfactual if and only if there exists a $\lambda \in \mathbb{R} \setminus \{0\}$ such that*

$$\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}_{CF}^* = b, \\
&\lambda \boldsymbol{a}^\top (\boldsymbol{x}_{CF}^* - \boldsymbol{x}_F) = \|\boldsymbol{x}_{CF}^* - \boldsymbol{x}_F\|_{N_1}, \\
&\|\lambda \boldsymbol{a}\|_{N_1}^* \leq 1.
\end{aligned}$$

To extract the original hyperplane, we need additional counterfactual queries. We note that a set of p linearly independent points $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}\}$ on the hyperplane, *i.e.*,

$$\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b = 0 \quad \forall i = 1, \ldots, p \tag{10}$$

is enough to extract $\boldsymbol{a}, b$. Thus, by finding p linearly independent counterfactual points, we can extract a hyperplane revealing the original classifier.

Remark 10. *In the case that one of our counterfactual queries returns $\mathbf{0}$ as a counterfactual for a certain point, we know that $b = 0$ since $\mathbf{0}$ lies on the hyperplane. Therefore, the solution space of $\boldsymbol{a}, b$ boils down to the solution space of $\boldsymbol{a}$, which is one-dimensional when we have $p-1$ linearly independent points $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p-1)}\}$. The system of equations described in* (10) *then simplifies to*

$$\boldsymbol{a}^\top \boldsymbol{x}^{(i)} = 0 \quad \forall i = 1, \ldots, p-1.$$

Since we only need the direction of $\boldsymbol{a}$ (any scaling leads to the same hyperplane), the latter equation system allows us to recover the classifier. Thus, if $\mathbf{0}$ is a counterfactual, then by finding $p-1$ linearly independent counterfactual vectors, the optimality conditions in Corollary 9 still ensure that we retrieve a hyperplane equivalent to the original classifier.

To obtain a set of p linearly independent points on the hyperplane, we strategically choose the factual instances to query. The following two lemmas are helping us with that.

Lemma 11. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance that does not lie on the hyperplane and $q_{CF}(\boldsymbol{x}_F) = \boldsymbol{x}_{CF}^*$ under an arbitrary norm $\|\cdot\|_{N_1}$. Then,*

$$\boldsymbol{x}_{CF}^* = \boldsymbol{x}_F + d_{\boldsymbol{x}_F} \boldsymbol{v}, \tag{11}$$

with $d_{\boldsymbol{x}_F} = \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F}{\|\boldsymbol{a}\|_{N_1}^*}$ is an optimal counterfactual if and only if $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$.

The lemma demonstrates the existence of a direction $\boldsymbol{v}$ such that, for any factual point, there is a corresponding counterfactual point obtained by perturbing the factual point in the direction of $\boldsymbol{v}$. Especially, the direction is independent of the factual instance. From the definition of the dual norm it follows that

$$\|\boldsymbol{a}\|_{N_1}^* = \max_{\|\boldsymbol{v}\|_{N_1} \leq 1} \boldsymbol{a}^\top \boldsymbol{v}, \tag{12}$$

and hence, we know that the direction $\boldsymbol{v}$ must be a maximizer of the latter problem.

If multiple optimal counterfactuals exist, one counterfactual query might perturb the original point in direction $\boldsymbol{v}^{(1)} := \boldsymbol{x}_{CF}^{(1)*} - \boldsymbol{x}_F^{(1)}$, while another counterfactual query might perturb the original

point in direction $\boldsymbol{v}^{(2)} = \boldsymbol{x}_{CF}^{(2)*} - \boldsymbol{x}_F^{(2)}$ with $\boldsymbol{v}^{(1)} \neq \boldsymbol{v}^{(2)}$. A consequence of Lemma 11 is that if such multiple optimal counterfactuals exist, we can always retrieve a counterfactual in the direction of $\boldsymbol{v}^{(1)}$. Concretely if for two factual points $\boldsymbol{x}_F^{(1)}, \boldsymbol{x}_F^{(2)} \in \mathcal{X}$ we have

$$\boldsymbol{x}_{CF}^{(1)*} = \boldsymbol{x}_F^{(1)} + \boldsymbol{v}^{(1)} \quad \text{and} \quad \boldsymbol{x}_{CF}^{(2)*} = \boldsymbol{x}_F^{(2)} + \boldsymbol{v}^{(2)},$$

where $\boldsymbol{v}^{(1)} \neq \boldsymbol{v}^{(2)}$ and $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)} \neq \mathbf{0}$. Then

$$\bar{\boldsymbol{x}}_{CF}^{(1)*} = \boldsymbol{x}_F^{(1)} + \frac{\|\boldsymbol{v}^{(1)}\|_{N_1}}{\|\boldsymbol{v}^{(2)}\|_{N_1}} \boldsymbol{v}^{(2)} \quad \text{and} \quad \bar{\boldsymbol{x}}_{CF}^{(2)*} = \boldsymbol{x}_F^{(2)} + \frac{\|\boldsymbol{v}^{(2)}\|_{N_1}}{\|\boldsymbol{v}^{(1)}\|_{N_1}} \boldsymbol{v}^{(1)}$$

are also optimal counterfactuals. This follows from Lemma 11 because if for one factual point there exists a counterfactual in the direction of $\boldsymbol{v}^{(1)}$, then for any point there exists a counterfactual in the direction of $\boldsymbol{v}^{(1)}$. If for another factual point, there exists a counterfactual in a different direction, we obtain the distance between the counterfactual and the factual. To obtain a counterfactual in the direction of $\boldsymbol{v}^{(1)}$, we only have to rescale the perturbation to match this distance between the factual and the counterfactual.

Example 12 (ℓ_1 and ℓ_∞ norm)**.** *For both the ℓ_1 and ℓ_∞ norms, a maximizer of* (12) *over the corresponding norm ball is attained at a vertex of the feasible region. Hence, for every factual instance $\boldsymbol{x}_F \in \mathcal{X}$ there exists an optimal counterfactual of the form $\boldsymbol{x}_{CF}^* = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$, where $\boldsymbol{v}$ is a vertex of the norm ball.*

For $N_1 = \ell_1$, this means that there exists an optimal counterfactual that modifies only one coordinate j_0 of $\boldsymbol{x}_F$. This becomes clear when choosing $\boldsymbol{v} = \operatorname{sgn}(a_{j_0})\, \boldsymbol{e}^{j_0}$, where $j_0 \in \arg\max_{j=1,\dots,p} |a_j|$. Then, $\|\boldsymbol{v}\|_1 = 1$ holds and $\boldsymbol{a}^\top \boldsymbol{v} = |a_{j_0}| = \|\boldsymbol{a}\|_\infty = \|\boldsymbol{a}\|_1^*$. However, the counterfactual query does not need to return a vertex solution when multiple optimal counterfactuals exist, i.e., when a higher-dimensional face of the ℓ_1-ball with center $\boldsymbol{x}_F$ and radius $\|\boldsymbol{x}_F - \boldsymbol{x}_{CF}^*\|_1$ intersects the hyperplane. Suppose that the query returns an optimal $\boldsymbol{x}_{CF}^* = \boldsymbol{x}_F + \sum_{i=1}^p d_i \boldsymbol{e}^i$ with at least two indices i_1, i_2 such that $d_{i_1}, d_{i_2} \neq 0$. Then, $\tilde{\boldsymbol{x}}_{CF} := \boldsymbol{x}_F + \left(\sum_{i=1}^p d_i\right)\boldsymbol{e}^{i_1}$ lies on the same face of the ℓ_1-ball and it has the same ℓ_1-distance to $\boldsymbol{x}_F$. Hence, $\tilde{\boldsymbol{x}}_{CF}$ is also an optimal counterfactual and it changes only one coordinate of the factual instance.

For $N_1 = \ell_\infty$ we can choose $\boldsymbol{v} = \sum_{i=1}^p \operatorname{sgn}(a_i)\boldsymbol{e}^i$. Then, $\|\boldsymbol{v}\|_\infty = 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \sum_{i=1}^p |a_i| = \|\boldsymbol{a}\|_1 = \|\boldsymbol{a}\|_\infty^*$, so there is an optimal counterfactual obtained by translating $\boldsymbol{x}_F$ towards a corner of the ℓ_∞-ball. Again, when a higher-dimensional face of the ℓ_∞-ball with center $\boldsymbol{x}_F$ and radius $\|\boldsymbol{x}_F - \boldsymbol{x}_{CF}^*\|_\infty$ intersects the hyperplane, the query may return a non-corner optimal point $\boldsymbol{x}_{CF}^* = \boldsymbol{x}_F + \sum_{i=1}^p d_i \boldsymbol{e}^i$ with at least two indices i_1, i_2 such that $|d_{i_1}| \neq |d_{i_2}|$. Let $j \in \arg\max_{i=1,\dots,p} |d_i|$, then $\tilde{\boldsymbol{x}}_{CF} := \boldsymbol{x}_F + d_j \sum_{i=1}^p \operatorname{sgn}(d_i)\boldsymbol{e}^i$ lies on the same face of the ℓ_∞-ball and it has the same ℓ_∞-distance to $\boldsymbol{x}_F$. Thus, $\tilde{\boldsymbol{x}}_{CF}$ is an optimal counterfactual obtained by translating $\boldsymbol{x}_F$ via a corner point of the ℓ_∞-ball.

Consequently, for both norms, there always exists an optimal counterfactual that moves the factual instance along a vertex direction of the respective norm ball, and this optimal counterfactual can be reconstructed from any optimal solution.

Lemma 13. *Let $\boldsymbol{v}$ be a vector from Lemma 11 and $V = \{\boldsymbol{v}, \boldsymbol{v}^2, \dots, \boldsymbol{v}^p\}$ be a basis of $\mathbb{R}^p$. If $\boldsymbol{v}_{CF}$ denotes the optimal counterfactual of $\boldsymbol{v}$ and $\boldsymbol{v}_{CF}^i$ the optimal counterfactual of $\boldsymbol{v}^i$ as given by* (11)*, then we have the following:*

(i) If $\boldsymbol{v}_{CF} \neq \mathbf{0}$, then the set of counterfactuals $\{\boldsymbol{v}_{CF}, \boldsymbol{v}_{CF}^2, \dots, \boldsymbol{v}_{CF}^p\}$ is linearly independent.

(ii) If $\boldsymbol{v}_{CF} = \mathbf{0}$, then the set of counterfactuals $\{\boldsymbol{v}_{CF}^2, \dots, \boldsymbol{v}_{CF}^p\}$ is linearly independent.

Lemmas 11 and 13 show that we can always construct a counterfactual of the form $\boldsymbol{x}_{CF} = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$, and if we have a basis V such that $\boldsymbol{v} \in V$, then querying counterfactuals for the whole basis V will lead to a linearly independent set of points that lie on the hyperplane. This allows us to recover all parameters $\boldsymbol{a}, b$.

Since we do not know $\boldsymbol{a}, b$ a priori, we also do not know the vector $\boldsymbol{v}$ from Lemma 11. However, after one counterfactual query for factual $\boldsymbol{x}_F$ such that $\boldsymbol{x}_F \neq \boldsymbol{x}_{CF}$ we can obtain $\boldsymbol{v}$ from Lemma 11 as

$$\boldsymbol{v} = \frac{1}{d_{\boldsymbol{x}_F}} \left(\boldsymbol{x}_{CF} - \boldsymbol{x}_F \right).$$

Afterwards, we can use the Gram-Schmidt process to create a basis V that contains $\boldsymbol{v}$. This procedure is described in Algorithm 1.

Algorithm 1 Extracting $\hat{\boldsymbol{a}}, \hat{b}$ using counterfactuals with non-differentiable norm $\| \cdot \|_{N_1}$

Input: Counterfactual query q_{CF}, norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$.
for $i = 1, \ldots, p$ **do**
  Query $\boldsymbol{x}_{CF}^{(i)*} \leftarrow q_{CF}(\boldsymbol{e}^i)$
  if $\boldsymbol{x}_{CF}^{(i)*} \neq \boldsymbol{e}^i$ **then**
    $\boldsymbol{v} \leftarrow \boldsymbol{x}_{CF}^{(i)*} - \boldsymbol{e}^i$
    break *for-loop*
  end if
end for
Create basis V such that $\boldsymbol{v} \in V$ using the Gram-Schmidt process
Query the points $\boldsymbol{v}_{CF}^i \leftarrow q_{CF}(\boldsymbol{v}^i)$ for all $\boldsymbol{v}^i \in V$
$\hat{\boldsymbol{a}}, b \leftarrow$ find solution to the set of linear equations (10)
return $\hat{\boldsymbol{a}}, \hat{b}$

Theorem 14. *Assuming $\boldsymbol{x}_{CF}^{(1)*} \neq \boldsymbol{e}^{(1)}$, Algorithm 1 uses $p+1$ counterfactual queries to return parameters $\hat{\boldsymbol{a}}, \hat{b}$ of a hyperplane, which is equivalent to the original hyperplane with parameters $\boldsymbol{a}, b$.*

Note that the assumption in Theorem 14 is not too restrictive since an arbitrary point as $\boldsymbol{e}^{(1)}$ is unlikely to lie on the hyperplane. Removing this assumption, it could be possible that all directions $\boldsymbol{e}^i$ need to be queried to find $\boldsymbol{v}$. However, then we would already have found p independent points on the hyperplane, so we can retrieve $\boldsymbol{a}, b$. When $\boldsymbol{v}$ is found after querying $p-1$ directions, then Algorithm 1 will query a newly created basis, hence in total $2p-1$ counterfactual queries are needed in the worst case.

Example 15. *We demonstrate the effectiveness of our approach using a simple two-dimensional example shown in Figure 4. Consider the hyperplane given by $2x_1 - x_2 = 3$, i.e., $\boldsymbol{a} = (2, -1), b = 3$. We examine a counterfactual mechanism with $N_1 = \ell_\infty$ as minimal edit. First, we ask a counterfactual query for the point $(3, 0)$, which yields its optimal counterfactual point $(2, 1)$. This gives us the direction of the counterfactuals, $(2, 1) - (3, 0) = (-1, 1)$. Next, we use a counterfactual query for the point $(-1, 1)$, which outputs $q_{CF}(-1, 1) = (1, -1)$. Note that we have already obtained a linearly independent set of equations to determine the hyperplane parameters:*

$$2a_1 + a_2 = b, \qquad a_1 - a_2 = b.$$

By setting $b = 1$, we obtain the solution $\boldsymbol{a} = \frac{1}{3}(2, 1)$. So, we found a hyperplane that is equivalent to the original hyperplane.

5 Robust Counterfactual Queries

Building upon the results of the counterfactual queries, we will first discuss in this section the number of robust counterfactual (RCF) queries needed to extract the original hyperplane. Additionally, we will show how the induced *'Yes'* and *'No'* regions are characterized when not enough queries are given.

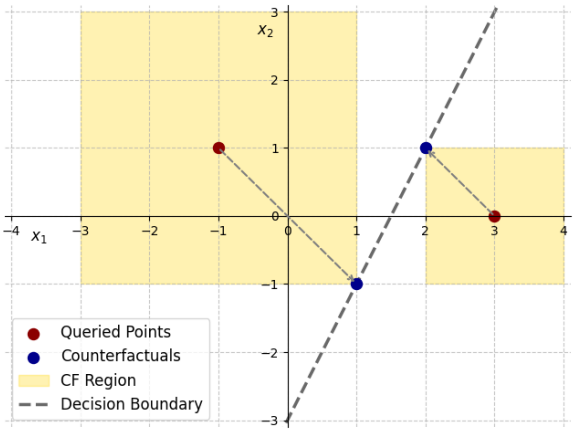


Figure 4: Example of extracting the hyperplane using counterfactual queries with $N_1 = \ell_\infty$.

5.1 Extracting the Hyperplane

In the following, we will show that the number of RCF queries needed to recover the linear classifier $h_{\boldsymbol{a},b}(\boldsymbol{x})$ depends on the norm of the robust counterfactual problem (2) but is independent of the norm used for the robustness set. The results show that the number of queries needed doubles compared to the classical counterfactuals, which indicates that in our setup, where the robustness radius ρ is known, robust counterfactuals provide an additional level of privacy compared to classical counterfactuals. We first show the following lemma, which is the robust counterpart of Lemma 7.

Lemma 16. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $\boldsymbol{x}^*_{RCF} \neq \boldsymbol{x}_F$ be a corresponding optimal robust counterfactual under an arbitrary norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$ and robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$ with $\rho > 0$. Then, there exists a scalar $\lambda^* \in \mathbb{R} \setminus \{0\}$ such that*

$$b = \boldsymbol{a}^\top \boldsymbol{x}^*_{RCF} + q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|^*_{N_2},$$
$$\lambda^* \boldsymbol{a} \in \partial f(\boldsymbol{x}^*_{RCF} - \boldsymbol{x}_F).$$

As for the classical counterfactuals, the lemma indicates that for differentiable norms, only one query is needed, while the situation with non-differentiable norms is more complicated.

5.1.1 Differentiable Norms

For differentiable norms, the subdifferential in Lemma 16 is a singleton, containing only the gradient of f. Hence, in this case the parameters of $\boldsymbol{a}$ can directly be extracted. The following theorem shows that we only need one counterfactual query to recover the hyperplane of the classifier.

Theorem 17. *Let q_{RCF} be a robust counterfactual mechanism using an arbitrary differentiable norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$ and robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$. One robust counterfactual query and one factual query for an arbitrary data point $\boldsymbol{x}_F \in \mathcal{X}$ are needed to extract the original model's parameters. In particular, for $q_{RCF}(\boldsymbol{x}_F) = \boldsymbol{x}^*_{RCF}$, let $\hat{\boldsymbol{a}} = \nabla f(\boldsymbol{x}^*_{RCF} - \boldsymbol{x}_F)$ and $\hat{b} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}^*_{RCF} + q_F(\boldsymbol{x}_F)\rho\|\hat{\boldsymbol{a}}\|^*_{N_2}$. Then, it holds that the hyperplane given by $\hat{\boldsymbol{a}}, \hat{b}$ is equivalent to the original hyperplane with parameters $\boldsymbol{a}, b$.*

Note that Theorem 8 finds an equivalent hyperplane as used by the original classifier, but it could be that the two different classification regions are interchanged. To obtain an equivalent classifier, it suffices to check the outcome of a single factual query, which is already done in Theorem 8. Recall that one counterfactual query sufficed for extracting the original model parameters, but an additional

factual query is required to determine the classification of the regions. Therefore, extracting the model parameters requires one more query when using robust counterfactuals, while finding an equivalent model requires the same number of queries.

5.1.2 Non-differentiable Norms

For non-differentiable norms, the subdifferential set in Lemma 16 can be larger than a singleton, so Theorem 17 no longer holds. In order to extract the hyperplane, we follow a similar procedure as for the counterfactual queries. The robust counterpart of Corollary 9 is derived in the following.

Corollary 18. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $\boldsymbol{q}_{RCF}$ be a robust counterfactual query under an arbitrary norm $\|\boldsymbol{x}\|_{N_1}$ and robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$. Then, $\boldsymbol{x}^*_{RCF} \neq \boldsymbol{x}_F$ is an optimal robust counterfactual if and only if there exists a $\lambda \in \mathbb{R} \setminus \{0\}$ such that*

$$\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}^*_{RCF} + q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|^*_{N_2} = b,\\
&\lambda \boldsymbol{a}^\top (\boldsymbol{x}^*_{RCF} - \boldsymbol{x}_F) = \|\boldsymbol{x}^*_{RCF} - \boldsymbol{x}_F\|_{N_1},\\
&\|\lambda \boldsymbol{a}\|^*_{N_1} \leq 1.
\end{aligned}$$

To retrieve the original hyperplane, we need additional robust counterfactual queries resulting in a system of equalities which has solution $\boldsymbol{a}, b$. Unlike counterfactuals, robust counterfactuals do not lie on the hyperplane. When we have a set of p linearly independent RCF points $\{\boldsymbol{x}^{(1)}_{RCF}, \ldots, \boldsymbol{x}^{(p)}_{RCF}\}$, we have the following system of equations

$$\boldsymbol{a}^\top \boldsymbol{x}^{(i)}_{RCF} - b + q_F(\boldsymbol{x}^{(i)}_F)\rho\|\boldsymbol{a}\|^*_{N_2} = 0 \quad \forall i = 1, \ldots, p, \tag{13}$$

which is nonlinear in $\boldsymbol{a}, b$ due to the term $\|\boldsymbol{a}\|^*_{N_2}$. When we know the classification of the original factual, we know $q_F(\boldsymbol{x}^{(i)}_F)$. After scaling we can also assume $\|\boldsymbol{a}\|^*_{N_2} = 1$ and (13) can be reformulated as

$$\begin{aligned}
\boldsymbol{a}^\top \boldsymbol{x}^{(i)}_{RCF} - b + q_F(\boldsymbol{x}^{(i)}_F)\rho = 0 \quad \forall i = 1, \ldots, p,\\
\|\boldsymbol{a}\|^*_{N_2} = 1.
\end{aligned} \tag{14}$$

The first p equations have a one-dimensional solution space since it is a linear system in dimension $p+1$ with p linearly independent equations. By solving the latter system and afterwards scaling the solution to $\|\boldsymbol{a}\|^*_{N_2} = 1$, we retrieve a hyperplane equivalent to the original classifier. Note that the latter discussion indicates that for each of the constructed factual points, we need to perform a factual query, which was not needed for the classical counterfactual case.

Remark 19. *In the case that $\mathbf{0}$ is a robust counterfactual for an arbitrary point $\boldsymbol{x} \in \mathcal{X}$, we obtain by the optimality conditions that $b = -q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|^*_{N_2}$. If we know the original classification of $\boldsymbol{x}$, the solution space of $\boldsymbol{a}, b$ boils down to the solution space of $\boldsymbol{a}$, which is one-dimensional when we have $p-1$ linearly independent points $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p-1)}\}$ such that*

$$\begin{aligned}
\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b + q_F(\boldsymbol{x}^{(i)}_F)\rho = 0 \quad \forall i = 1, \ldots, p-1,\\
\|\boldsymbol{a}\|^*_{N_2} = 1.
\end{aligned}$$

Thus, if $\mathbf{0}$ is a robust counterfactual, then, by finding $p-1$ linearly independent counterfactual vectors, the optimality conditions in Lemma 18 still ensure that we retrieve a hyperplane equivalent to the original classifier.

The latter discussion indicates that we have to find a set of p linearly independent RCF points. To achieve this, we have to smartly choose the factual instances $\boldsymbol{x}^{(i)}_F$. To approach this we show the following lemma.

Lemma 20. *Let $\boldsymbol{x}_F \in \mathcal{X}$ be an arbitrary factual instance and $q_{RCF}(\boldsymbol{x}_F) = \boldsymbol{x}^*_{RCF}$ under an arbitrary norm $\|\cdot\|_{N_1}$ and robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$. Then,*

$$\boldsymbol{x}^*_{RCF} = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v},$$

with $d_{\boldsymbol{x}_F} = \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|^*_{N_2}}{\|\boldsymbol{a}\|^*_{N_1}}$ is an optimal robust counterfactual if and only if $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$.

The lemma shows, similar to the results for counterfactuals, that there exists a direction $\boldsymbol{v}$ such that for any factual point, there exists a robust counterfactual which is the factual point perturbed into this direction $\boldsymbol{v}$. Especially, the direction is independent of the factual instance. From the definition of the dual norm it follows that

$$\|\boldsymbol{a}\|^*_{N_1} = \max_{\|\boldsymbol{v}\|_{N_1} \leq 1} \boldsymbol{a}^\top \boldsymbol{v},$$

and hence, we know that the direction $\boldsymbol{v}$ must be a maximizer of the latter problem. Especially, the direction $\boldsymbol{v}$ is the same as for classical counterfactuals.

By combining Lemma 13 with Lemma 20, we can apply the same algorithm used for counterfactuals to recover the hyperplane using robust counterfactuals when norm-1 is non-differentiable. Specifically, in Algorithm 1, we replace q_{CF} with both q_{RCF} and q_F. Consequently, we need $p+1$ robust counterfactual queries and $p+1$ factual queries to solve the system of equations (14) and obtain hyperplane parameters $\hat{\boldsymbol{a}}, \hat{b}$ that are equivalent to the original hyperplane parameters $\boldsymbol{a}, b$.

Corollary 21. *Let q_{RCF} be a robust counterfactual mechanism using an arbitrary non-differentiable norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$ and robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$. Then, with only $p+1$ robust counterfactual queries and $p+1$ factual queries, we can recover hyperplane parameters $\hat{\boldsymbol{a}}, \hat{b}$ which are equivalent to the original hyperplane parameters $\boldsymbol{a}, b$.*

Example 22. *We demonstrate our approach using a simple two-dimensional example shown in Fig. 5. Consider the hyperplane given by $2x_1 - x_2 = 3$, i.e., $\boldsymbol{a} = (2, -1), b = 3$. We examine a counterfactual mechanism with $N_1 = \ell_\infty$ as minimal edit and robustness set given by $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq 1\}$ with $N_2 = \ell_1$. First, we ask a factual query for $(3, 0)$, which outputs 'No'. A robust counterfactual query for the point $(3, 0)$ yields $q_{RCF}(3, 0) = \frac{1}{3}(4, 5)$. This gives us the same direction as for the counterfactuals, $(-1, 1)$. Next, a factual query $(-1, 1)$ yields 'Yes' and a robust counterfactual query for the point $(-1, 1)$ outputs $q_{RCF}(-1, 1) = \frac{1}{3}(5, -5)$. Note that we have obtained the following equations:*

$$\frac{1}{3}(4a_1 + 5a_2) - b - \|\boldsymbol{a}\|_\infty = 0 \quad \text{and} \quad \frac{1}{3}(5a_1 - 5a_2) - b + \|\boldsymbol{a}\|_\infty = 0.$$

By setting $\|\boldsymbol{a}\|_\infty = 1$, we obtain the solution $b = \frac{3}{2}a_1, a_2 = \frac{a_1}{10} - \frac{3}{5}$. Moreover, since we enforce $\|\boldsymbol{a}\|_\infty = 1$ we have $\max\{|a_1|, |\frac{a_1}{10} - \frac{3}{5}|\} = 1$ implying that $a_1 = -1$ or $a_1 = 1$. For $a_1 = -1$, we have the hyperplane $\boldsymbol{a} = (-1, -\frac{7}{10}), b = -\frac{3}{2}$. Then, for point $\boldsymbol{x} = \frac{1}{3}(4, 5)$, we would have $\boldsymbol{a}^\top \boldsymbol{x} - b = -\frac{4}{3} + \frac{3}{2} = \frac{1}{9} > 0$ which contradicts with $\boldsymbol{x}$ being a robust counterfactual of $(3, 0)$. Therefore, we conclude $a_1 = 1$ giving us the hyperplane $\boldsymbol{a} = (1, -\frac{1}{2}), b = \frac{3}{2}$, which is an equivalent hyperplane to the original one.

5.2 Classification Regions

In the previous section, we used Theorem 17 to conclude that if $\|\cdot\|_{N_1}$ is differentiable, then the full hyperplane is extracted using one factual and one counterfactual query. Hence, we exactly know the full classification regions. However, for non-differentiable norms, this is not the case.

Now, suppose that, next to a set of arbitrary factuals, we also have an arbitrary set of robust counterfactuals (RCFs). This means that for points $\boldsymbol{x}^{(j)}$ from an index subset $j \in J_0 \cup J_1 \subseteq I_0 \cup I_1$

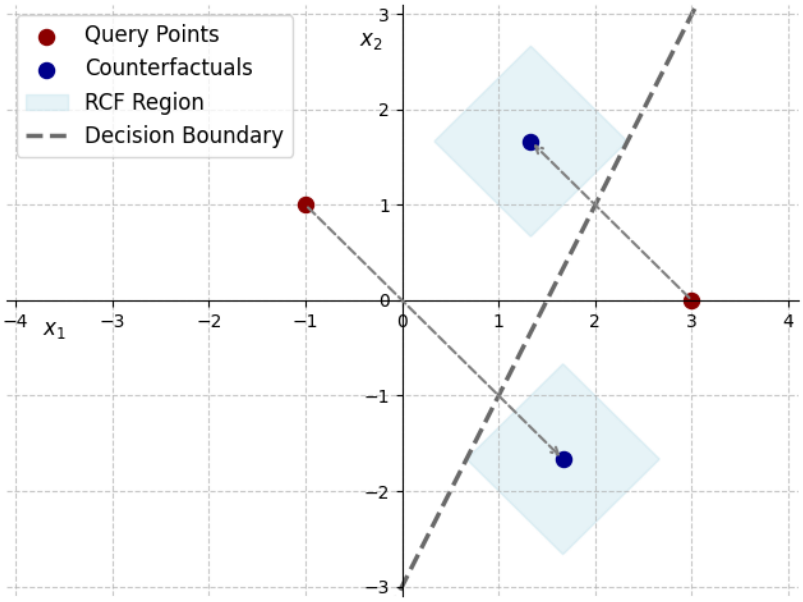


Figure 5: Example of extracting the hyperplane using robust counterfactual queries with $N_1 = \ell_\infty$ and robustness set given by $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq 1\}$ with $N_2 = \ell_1$.

we get a RCF $\boldsymbol{x}_{RCF}^{(j)}$. Then, we know that every point $\boldsymbol{z}$ in the robust region $\mathcal{S}$ around the robust counterfactual will be classified differently from the corresponding factual. Concretely,

$$\boldsymbol{a}^\top \boldsymbol{z} - b \geq 0 \quad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}_{RCF}^{(j)}\|_{N_2} \leq \rho \qquad \forall j \in J_0, \tag{15}$$

$$\boldsymbol{a}^\top \boldsymbol{z} - b \leq 0 \quad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}_{RCF}^{(j)}\|_{N_2} \leq \rho, \qquad \forall j \in J_1. \tag{16}$$

Besides, Lemma 16 tells us $\boldsymbol{a}$ is a subgradient of norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_{N_1}$ at $(\boldsymbol{x}_{RCF}^{(j)} - \boldsymbol{x}^{(j)})$. That is,

$$\boldsymbol{a} \in \partial f(\boldsymbol{x}_{RCF}^{(j)} - \boldsymbol{x}^{(j)}) \qquad \forall j \in J. \tag{17}$$

With this extra information, we can characterize the uncertainty set of $\boldsymbol{a}, b$ as $\mathcal{U}_{\boldsymbol{a},b}^{RCF}$ as given by the following set of constraints:

$$\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b \leq 0 && \forall i \in I_0, \\
&\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b \geq 0 && \forall i \in I_1, \\
&\boldsymbol{a}^\top \boldsymbol{z} - b \geq 0 \quad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}_{RCF}^{(j)}\|_{N_2} \leq \rho && \forall j \in J_0, \\
&\boldsymbol{a}^\top \boldsymbol{z} - b \leq 0 \quad \forall \boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}_{RCF}^{(j)}\|_{N_2} \leq \rho, && \forall j \in J_1, \\
&\boldsymbol{a} \in \partial f(\boldsymbol{x}_{RCF}^{(j)} - \boldsymbol{x}^{(j)}) && \forall j \in J.
\end{aligned}$$

We remark that for norm-1 equal to ℓ_1 or ℓ_∞, constraint (17) can be linearized as the subgradient of these norms can be modeled using linear constraints. The rest of the uncertainty set is of a similar structure to $\mathcal{U}_{\boldsymbol{a},b}^{CF}$ as presented for the counterfactual queries. Therefore, using the same techniques as presented in Section 4.1, a computationally tractable characterization of the *'Yes'* and *'No'* regions can be derived.

However, we know more information about $\boldsymbol{a}, b$. We know that the robust region $\mathcal{S}$ will touch the hyperplane since otherwise a closer robust counterfactual would exist. We generally cannot characterize where the robust region around the robust counterfactual touches the hyperplane without knowing

the model parameters. This is a key difference with general counterfactuals, where we know that the counterfactual point lies on the hyperplane. Using this information, we can rewrite constraints (15) and (16). To this end, we first rewrite constraint (15) for the worst case scenario to get an equivalent constraint

$$\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^{(j)} - b - \rho\|\boldsymbol{a}\|_{N_2}^* \geq 0 \quad \forall j \in J_0.$$

Moreover, since we also know $\mathcal{S}$ touches the hyperplane, equality should hold for this constraint, resulting in the first optimality condition of Lemma 16. Similarly, constraint (16) can be replaced by the following stronger constraint:

$$\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^{(j)} - b + \rho\|\boldsymbol{a}\|_{N_2}^* = 0 \quad \forall j \in J_1.$$

Therefore, a stricter uncertainty set of $\boldsymbol{a}, b$ denoted by $\overline{\mathcal{U}}_{\boldsymbol{a},b}^{RCF}$ can be described using the following constraints

$$\begin{aligned}
\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b &\leq 0 && \forall i \in I_0,\\
\boldsymbol{a}^\top \boldsymbol{x}^{(i)} - b &\geq 0 && \forall i \in I_1,\\
\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^{(j)} - b - \rho\|\boldsymbol{a}\|_{N_2}^* &= 0 && \forall j \in J_0\\
\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^{(j)} - b + \rho\|\boldsymbol{a}\|_{N_2}^* &= 0 && \forall j \in J_1,\\
\boldsymbol{a} &\in \partial f(\boldsymbol{x}_{RCF}^{(j)} - \boldsymbol{x}^{(j)}) && \forall j \in J.
\end{aligned}$$

Unfortunately, these equality constraints in the uncertainty set $\overline{\mathcal{U}}_{\boldsymbol{a},b}^{RCF}$ make the inner optimization problems for the *'Yes'* and *'No'* regions,

$$\mathcal{X}_{\text{'Yes'}} = \{\boldsymbol{x} \mid \min_{\boldsymbol{a},b \in \overline{\mathcal{U}}_{\boldsymbol{a},b}^{RCF}} \boldsymbol{a}^\top \boldsymbol{x} - b \geq 0\} \qquad \text{and} \qquad \mathcal{X}_{\text{'No'}} = \{\boldsymbol{x} \mid \max_{\boldsymbol{a},b \in \overline{\mathcal{U}}_{\boldsymbol{a},b}^{RCF}} \boldsymbol{a}^\top \boldsymbol{x} - b \leq 0\}$$

intractable due to the non-linearity of $\|\cdot\|_{N_2}^*$. In fact, the reformulation in the previous sections relied on a duality argument that cannot be applied in the latter situation.

Using the abovementioned uncertainty set, $\overline{\mathcal{U}}_{\boldsymbol{a},b}^{RCF}$, we visualize the classification regions when there is one data point classified as *'No'* and a corresponding robust counterfactual, *i.e.*, $|I| = |J| = 1$. We examined nine cases, combining norm-1 and norm-2 pairs from $\{\ell_1, \ell_2, \ell_\infty\}$. The results are depicted in Figure 6, where we see that for norm-1 non-differentiable, *i.e.*, ℓ_1, ℓ_∞, there are areas where we cannot conclude the classification.

Two notable insights result in conditions on $\boldsymbol{a}, b$, which can be added to $\mathcal{U}_{\boldsymbol{a},b}^{RCF}$ such that the dualization for the inner optimization problems for the *'Yes'* and *'No'* region remains tractable.

First, we see in Figure 6 that for norm-1 non-differentiable, the *'No'* region is a translated pointed cone. The vertex point of this cone is not always equal to the original factual queried data point. In the following, we will characterize this point $\bar{\boldsymbol{x}}$. This results in an extra data point for which we know the classification, which reduces the uncertainty in $\boldsymbol{a}, b$.

Lemma 23. *Let $\boldsymbol{x}_F$ be a data point and let $\boldsymbol{x}_{RCF}^*$ be an optimal robust counterfactual under $\|\cdot\|_{N_1}$-distance and with robustness set $\mathcal{S} = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_{N_2} \leq \rho\}$. Furthermore, assume that $\|\boldsymbol{a}\|_{N_2}^* \leq C\|\boldsymbol{a}\|_{N_1}^*$ for all $\boldsymbol{a} \in \mathbb{R}^p$. Then, for $\boldsymbol{v} = (\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F)/\|\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F\|_{N_1}$ the point $\bar{\boldsymbol{x}} := \boldsymbol{x}_{RCF} - d\boldsymbol{v}$ has the same classification as $\boldsymbol{x}_F$ if $d \geq \rho C$.*

We can conclude from the latter lemma the following special cases.

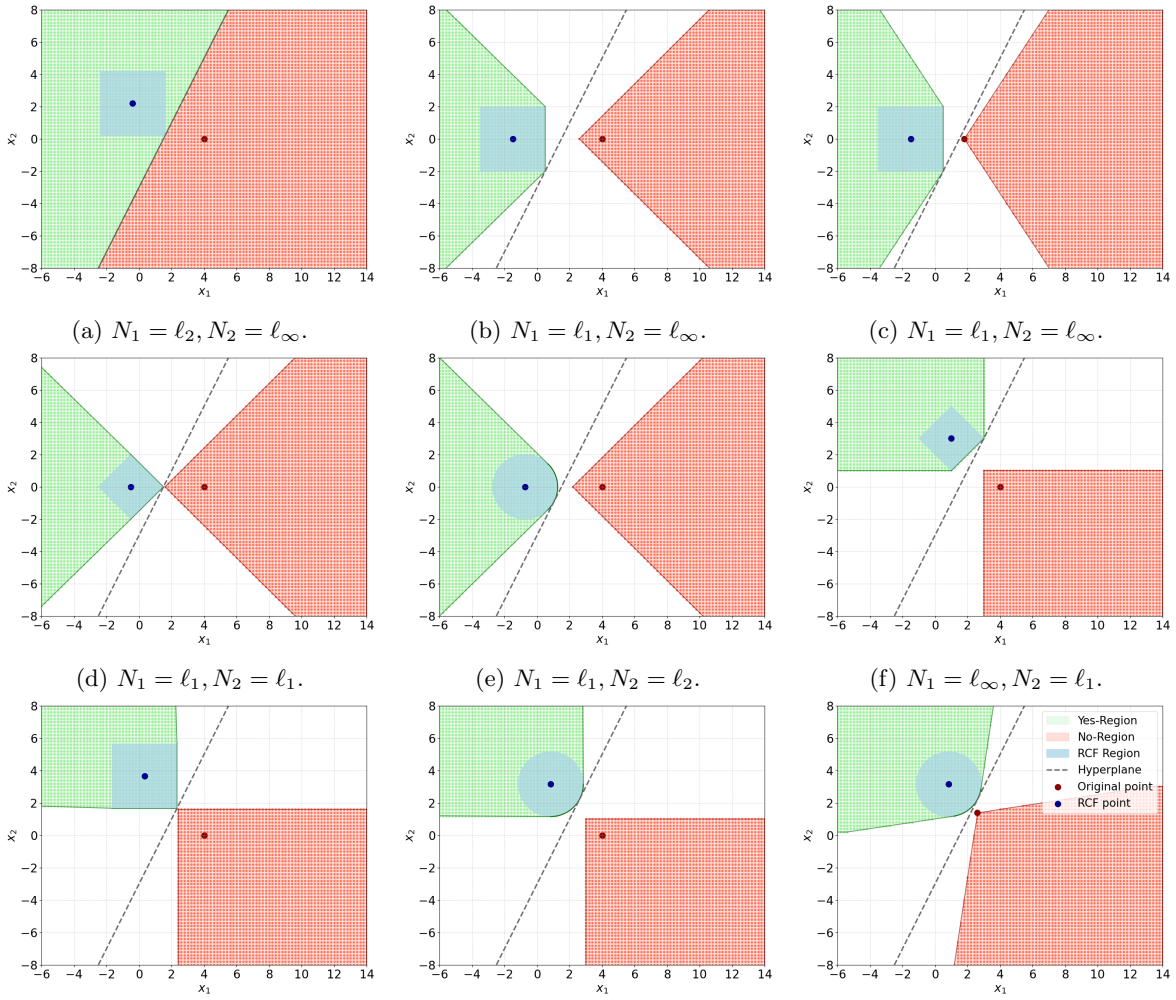
Corollary 24. *Under the same assumptions as in Lemma 23 the following results hold:*

(i) For $N_1 = \ell_1$ and $N_2 = \ell_q$ with $q \geq 1$, the point $\bar{\boldsymbol{x}}$ has the same classification as $\boldsymbol{x}_F$ when $d \geq \rho p^{1-1/q}$.

(ii) For $N_1 = \ell_\infty$ and $N_2 = \ell_q$ with $q \geq 1$, the point $\bar{\boldsymbol{x}}$ has the same classification as $\boldsymbol{x}_F$ when $d \geq \rho$.

(iii) If $N_1 = N_2$, then $\boldsymbol{x}_{\mathcal{S}} = \boldsymbol{x}^*_{RCF} - \rho \boldsymbol{v}$ lies on the hyperplane.

The latter corollary shows that in the case when $N_1 = N_2$, see, *e.g.*, Figure 6d and Figure 6g, it is possible to characterize an exact point $\boldsymbol{x}_{\mathcal{S}}$, where the robust region $\mathcal{S}$ around the robust counterfactual touches the hyperplane. Using this point, constraints (15) and (16) can be reformulated to inequalities stating the robustness set to be on the desired side of the hyperplane, while adding constraint $\boldsymbol{a}^\top \boldsymbol{x}_{\mathcal{S}} - b = 0$.



(a) $N_1 = \ell_2, N_2 = \ell_\infty$. (b) $N_1 = \ell_1, N_2 = \ell_\infty$. (c) $N_1 = \ell_1, N_2 = \ell_\infty$.

(d) $N_1 = \ell_1, N_2 = \ell_1$. (e) $N_1 = \ell_1, N_2 = \ell_2$. (f) $N_1 = \ell_\infty, N_2 = \ell_1$.

(g) $N_1 = \ell_\infty, N_2 = \ell_\infty$. (h) $N_1 = \ell_\infty, N_2 = \ell_2$. (i) $N_1 = \ell_\infty, N_2 = \ell_2$.

Figure 6: Examples of the classification regions given one *'No'* factual and corresponding robust counterfactual for different norms.

6 Conclusion

In this work, we presented how factual, counterfactual, and robust counterfactual queries can be used to extract information of linear classifiers in p-dimensional feature space. We provided computationally tractable characterizations of the *'Yes'* and *'No'* regions given an arbitrary set of factual, counterfactual, or robust counterfactual queried points. This allows us to efficiently check which points will be

classified as *'Yes'* and which as *'No'* without querying the model again. For factual queries, we have demonstrated that the classification regions extend beyond the convex hull of the factual data points. When we additionally consider counterfactuals, we gain more information about the classification regions. In particular, we have proven that only one counterfactual can extract the full hyperplane when the *minimal edit* distance function is differentiable. When this distance function is non-differentiable, $p + 1$ counterfactual queries suffice to retrieve the full hyperplane. We have seen that these results can be generalized for robust counterfactuals. We show that deriving tractable formulations for the *'Yes'* and *'No'* regions is more difficult than for classical counterfactuals but can be done for certain special cases. We have also shown that one pair of factual and robust counterfactual queries suffices to extract the original hyperplane when the *minimal edit* distance function is differentiable. When this distance function is non-differentiable, $p + 1$ pairs of factual and robust counterfactual queries are needed. Notably, the norm used to define the robust region does not affect the number of queries needed to extract the model, but can affect the information gained on the classification regions. In summary, our results show that to increase privacy of a linear model, using non-differentiable norms for the distance of the counterfactuals is beneficial. Furthermore, providing robust counterfactuals adds an additional layer of privacy, since for recovery of the model parameters additional factual queries are needed.

Our results naturally come with limitations. Since we assume unconstrained factual and counterfactual queries in $\mathbb{R}^p$, the results presented in this paper are not evidently generalizable to non-continuous data, *e.g.*, categorical or binary. In practice, some features may be immutable, leading to constrained counterfactuals, which we do not address. Future work includes extending our approach to linear models with kernels and to other non-linear models, such as classifiers using quadratic polynomials. It could also be applied to extract classification regions and parameters of decision tree models, which rely on separating hyperplanes. Furthermore, our work relies on the assumption that the queried (robust) counterfactuals are exact, *i.e.*, optimal solutions of the corresponding optimization problems. In practice, often heuristic methods are used, returning non-optimal (robust) counterfactuals. Extending our work for this situation could be of practical interest. Finally, another direction is to develop defense mechanisms against the model-extraction techniques proposed in this work.

References

Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

Basel Committee on Banking Supervision. The Basel Committee – Overview, 2026. URL `https://www.bis.org/bcbs/index.htm`. Last access: 1 February 2026.

Sjoerd Berning, Vincent Dunning, Dayana Spagnuelo, Thijs Veugen, and Jasper Van Der Waa. The trade-off between privacy & quality for counterfactual explanations. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pages 1–9, 2024.

Dimitris Bertsimas and Dick den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas LLC, 2022. ISBN 978-1-7337885-2-6.

Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 175–199. IEEE, 2023.

Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1309–1326, 2020.

Pasan Dissanayake and Sanghamitra Dutta. Model reconstruction using counterfactual explanations: A perspective from polytope theory. *Advances in Neural Information Processing Systems*, 37:83397–83429, 2024.

European Union. Art. 22 GDPR – Automated individual decision-making, including profiling, 2016. URL `https://gdpr.eu/article-22-automated-individual-decision-making/`. Last access: 1 February 2026.

Federal Reserve System. Sr 11-7: Guidance on Model Risk Management. Technical report, Board of Governors of the Federal Reserve System, Division of Banking Supervision and Regulation, August 2011. URL `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm`. Last access: 1 February 2026.

Julien Ferry, Ricardo Fukasawa, Timothée Pascal, and Thibaut Vidal. Trained random forests completely reveal your dataset. *arXiv preprint arXiv:2402.19232*, 2024.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Sofie Goethals, Kenneth Sörensen, and David Martens. The privacy issue of counterfactual explanations: explanation linkage attacks. *ACM Transactions on Intelligent Systems and Technology*, 14 (5):1–24, 2023.

Awa Khouna, Julien Ferry, and Thibaut Vidal. From counterfactuals to trees: Competitive analysis of model extraction attacks. *arXiv preprint arXiv:2502.05325*, 2025.

Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Siscovery in Data Mining*, pages 641–647, 2005.

Donato Maragno, Jannis Kurtz, Tabea E Röber, Rob Goedhart, Ş İlker Birbil, and Dick den Hertog. Finding regions of counterfactual explanations via robust optimization. *INFORMS Journal on Computing*, 36(5):1316–1334, 2024.

Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024.

Abdullah Caglar Oksuz, Anisa Halimi, and Erman Ayday. Autolycus: Exploiting explainable artificial intelligence (XAI) for model extraction attacks against interpretable models. *Proceedings on Privacy Enhancing Technologies*, 2024(4):684–699, October 2024. ISSN 2299-0984. doi: 10.56553/popets-2024-0137. URL `http://dx.doi.org/10.56553/popets-2024-0137`.

Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 865–872, 2020.

Robert Nikolai Reith, Thomas Schneider, and Oleksandr Tkachenko. Efficiently stealing your machine learning models. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*, pages 198–210, 2019.

Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1997.

Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841, 2017.

Yongjie Wang, Hangwei Qian, and Chunyan Miao. Dualcf: Efficient model extraction attack from counterfactual explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1318–1329, 2022.

A Proofs

A.1 Factuals

Theorem 5. When we examine the inner optimization problem of (4), we can use dualization with respect to $\boldsymbol{a}, b$ to get an equivalent system of equations:

$$\begin{array}{lll} \max & \boldsymbol{a}^\top \boldsymbol{x} - b & \\ s.t. & \boldsymbol{a}, b \in \mathcal{U}^F_{\boldsymbol{a},b}. & \end{array} \iff \begin{array}{lll} \max & (\boldsymbol{x}, -1)^\top \boldsymbol{a}, b & \\ s.t. & (\boldsymbol{x}^{(i)}, -1)^\top \boldsymbol{a}, b \leq 0 & \forall i \in I_0, \\ & (-\boldsymbol{x}^{(i)}, 1)^\top \boldsymbol{a}, b \leq 0 & \forall i \in I_i. \end{array}$$

$$\iff \begin{array}{lll} \min & \mathbf{0} & \\ s.t. & \sum_{i \in I_0} u_i \boldsymbol{x}^{(i)} - \sum_{i \in I_1} u_i \boldsymbol{x}^{(i)} = \boldsymbol{x}, & \\ & -\sum_{i \in I_0} u_i + \sum_{i \in I_1} u_i = -1, & \\ & \boldsymbol{u} \geq \mathbf{0}. & \end{array}$$

Hence, the *'No'* region can be written as the following polyhedral set:

$$\mathcal{X}_{\text{'No'}} = \left\{ \boldsymbol{x} \mid \exists \boldsymbol{u} \ : \ \sum_{i \in I_0} u_i - \sum_{i \in I_1} u_i = 1, \ \sum_{i \in I_0} \boldsymbol{x}^{(i)} u_i - \sum_{i \in I_1} \boldsymbol{x}^{(i)} u_i = \boldsymbol{x}, \ \boldsymbol{u} \geq \mathbf{0} \right\}.$$

Similarly, using dualization the *'Yes'* region can be described as

$$\mathcal{X}_{\text{'Yes'}} = \left\{ \boldsymbol{x} \mid \exists \boldsymbol{u} \ : \ \sum_{i \in I_1} u_i - \sum_{i \in I_0} u_i = 1, \ \sum_{i \in I_1} \boldsymbol{x}^{(i)} u_i - \sum_{i \in I_0} \boldsymbol{x}^{(i)} u_i = \boldsymbol{x}, \ \boldsymbol{u} \geq \mathbf{0} \right\}.$$

□

A.2 Counterfactuals

Theorem 6. To characterize the *'No'* region, we need to check when the inner optimization problem as described in equation 4 is at most 0. Using the uncertainty set of $\boldsymbol{a}, b$ this optimization problem becomes

$$\begin{array}{lrll} \max\limits_{\boldsymbol{a},b} & \boldsymbol{a}, b^\top (\boldsymbol{x}, -1) & & \\ s.t. & \boldsymbol{a}, b^\top (\boldsymbol{x}^{(i)}, -1) & \leq 0 & \forall i \in I_0, \\ & \boldsymbol{a}, b^\top (-\boldsymbol{x}^{(i)}, 1) & \leq 0 & \forall i \in I_1, \\ & \boldsymbol{a}, b^\top \boldsymbol{z} & \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{U}_j & \forall j \in J_0, \\ & \boldsymbol{a}, b^\top \boldsymbol{z} & \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{U}_j & \forall j \in J_1, \\ & \boldsymbol{a}, b^\top (\boldsymbol{x}^{(j)}_{CF}, -1) & = 0 & \forall j \in J, \end{array}$$

where the uncertainty sets for $j \in J_0$ are given by

$$\mathcal{U}_j = \{ \boldsymbol{z} := (\boldsymbol{z}_a, z_b) \mid \| \boldsymbol{z}_a - \boldsymbol{x}^{(j)} \|_{N_1} \leq \rho_j, \ z_b = -1 \}$$

and for $j \in J_1$ by

$$\mathcal{U}_j = \{ \boldsymbol{z} := (\boldsymbol{z}_a, z_b) \mid \| \boldsymbol{z}_a + \boldsymbol{x}^{(j)} \|_{N_1} \leq \rho_j, \ z_b = 1 \}.$$

We clearly see that for each $j \in J$ the uncertainty set can be written as $\mathcal{U}_j = \{ \boldsymbol{z} \mid f_{jk}(\boldsymbol{z}) \leq 0, \forall k \in K_j \}$, where f_{jk} is a convex function and K_j the set of indices of constraints that defines uncertainty set $\mathcal{U}_j$. Besides, we see that the uncertainty sets are bounded. Using the dualization results, primal worst is dual best, as presented in Bertsimas and den Hertog [2022, equation (2.45)], we can conclude this optimization problem is equivalent to:

$$\begin{array}{lll} \min\limits_{\boldsymbol{t},\boldsymbol{u},\boldsymbol{v},\boldsymbol{z}} & 0 & \\ s.t. & \sum\limits_{i \in I_0} t_i (\boldsymbol{x}^{(i)}, -1) + \sum\limits_{i \in I_1} t_i (-\boldsymbol{x}^{(i)}, 1) + \sum\limits_{j \in J} v_j (\boldsymbol{x}^{(j)}_{CF}, -1) + \sum\limits_{j \in J} u_j \boldsymbol{z}^{(j)} & = (\boldsymbol{x}, -1), \\ & \boldsymbol{z}^{(j)} \in \mathcal{U}_j \ \ \forall j \in J, & \\ & \boldsymbol{t}, \boldsymbol{u} \in \mathbb{R}^{|I|}_{\geq 0}, \ \boldsymbol{v} \in \mathbb{R}^{|J|}, \ \boldsymbol{z}^{(j)} \in \mathbb{R}^{p+1} \ \ \forall j \in J. & \end{array}$$

Let $\boldsymbol{y}^j = (\boldsymbol{y}_a^j, y_b^j) = u_j \boldsymbol{z}^{(j)}$. Then, following Bertsimas and den Hertog [2022, Theorem 2.2], we write it as the following conic quadratic problem:

$$
\begin{array}{lll}
\min\limits_{\boldsymbol{t},\boldsymbol{u},\boldsymbol{v},\boldsymbol{y}} & 0 & \\
s.t. & \sum_{i\in I_0} t_i(\boldsymbol{x}^{(i)}, -1) + \sum_{i\in I_1} t_i(-\boldsymbol{x}^{(i)}, 1), & \\
& + \sum_{j\in J} v_j(\boldsymbol{x}_{CF}^{(j)}, -1) + \sum_{j\in J}(\boldsymbol{y}_a^{(j)}, y_b^{(j)}) = (\boldsymbol{x}, -1), & \\
& u_j(\|\boldsymbol{y}_a^{(j)}/u_j - \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 & \forall j \in J_0, \\
& u_j(y_b^{(j)}/u_j + 1) = 0 & \forall j \in J_0, \\
& u_j(\|\boldsymbol{y}_a^{(j)}/u_j + \boldsymbol{x}^{(j)}\|_{N_1} - \rho_j) \leq 0 & \forall j \in J_1, \\
& u_j(y_b^{(j)}/u_j - 1) = 0 & \forall j \in J_1, \\
& \boldsymbol{t} \in \mathbb{R}^{|I|}_{\geq 0},\ \boldsymbol{u} \in \mathbb{R}^{|J|}_{\geq 0},\ \boldsymbol{v} \in \mathbb{R}^{|J|},\ \boldsymbol{y}^{(j)} \in \mathbb{R}^{p+1}\ \ \forall j \in J. &
\end{array}
$$

Since the objective of the latter optimization problem is constant 0, finding out whether $\boldsymbol{x}$ will be classified as a *'No'* can be determined by checking if there exist $\boldsymbol{t}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{y}$ that meet the constraints. The proof for the *'Yes'* region follows a similar argumentation. □

Lemma 7. Without loss of generality, we may assume that $\boldsymbol{x}_F$ is classified as *'No'*. Given the classifier $h_{\boldsymbol{a},b}$, a potential counterfactual point must lie in the half-space $\boldsymbol{a}^\top \boldsymbol{x} \geq b$. Since we are minimizing the distance to $\boldsymbol{x}_F$, an optimal counterfactual actually lies on the boundary of the half-space $\boldsymbol{a}^\top \boldsymbol{x} = b$. Hence, the optimization problem (1) can be reformulated as

$$
\begin{aligned}
&\min_{\boldsymbol{x}_{CF}} \|\boldsymbol{x}_{CF} - \boldsymbol{x}_F\|_{N_1} \\
&s.t. \quad \boldsymbol{a}^\top \boldsymbol{x}_{CF} = b.
\end{aligned}
$$

The latter problem has a convex objective function and one linear equality constraint. Hence, we know that every optimal solution $\boldsymbol{x}^*$ must fulfill the KKT-conditions

$$
\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}^* = b, \\
&\boldsymbol{0} \in \partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*),
\end{aligned}
$$

where λ^* is a dual optimal solution and $\mathcal{L}$ is the Lagrangean dual function

$$
\mathcal{L}(\boldsymbol{x}, \lambda) = \|\boldsymbol{x} - \boldsymbol{x}_F\|_{N_1} - \lambda(\boldsymbol{a}^\top \boldsymbol{x} - b).
$$

We have $\partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda) = \partial f(\boldsymbol{x} - \boldsymbol{x}_F) - \lambda \boldsymbol{a}$ which proves the result. Note, that since $\boldsymbol{x}_F \neq \boldsymbol{x}_{CF}^*$ we know that $\boldsymbol{0} \notin \partial f(\boldsymbol{x}_{CF}^* - \boldsymbol{x}_F)$, and hence, $\lambda^* \neq 0$ must hold. □

Theorem 8. Without loss of generality, we assume that $\boldsymbol{x}_F$ is classified as *'No'*. We apply Lemma 7 to the case where the norm is differentiable. Then, there exists an $\lambda^* \neq 0$ such that

$$
\begin{aligned}
&\lambda^* \boldsymbol{a} = \nabla f(\boldsymbol{x}_{CF}^* - \boldsymbol{x}_F), \\
&b = \boldsymbol{a}^\top \boldsymbol{x}_{CF}^*.
\end{aligned}
$$

By defining $\hat{\boldsymbol{a}} = \nabla f(\boldsymbol{x}_{CF}^* - \boldsymbol{x}_F)$ and $\hat{b} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_{CF}^*$, we have $\lambda^* \boldsymbol{a} = \hat{\boldsymbol{a}}$ and $\lambda^* b = \lambda^* \boldsymbol{a}^\top \boldsymbol{x}_{CF}^* = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_{CF}^* = \hat{b}$. Hence, we $\frac{1}{\lambda^*}\boldsymbol{a}, b = \boldsymbol{a}, b$ for a $\lambda^* \neq 0$ and the hyperplane given by $\boldsymbol{a}, b$ is equivalent to the original hyperplane with parameters $\boldsymbol{a}, b$ which proves the result. □

Lemma 11. We note that $\boldsymbol{a} \neq \boldsymbol{0}$, hence $\|\boldsymbol{a}\|^* \neq \boldsymbol{0}$ and $d_{\boldsymbol{x}_F}$ is well defined. Besides, since we consider points $\boldsymbol{x}_F$ that do not lie on the hyperplane, we know $d_{\boldsymbol{x}_F} > 0$. Moreover, we have $\|\boldsymbol{a}\|_{N_1}^* := \sup_{\|\boldsymbol{x}\|_{N_1} \leq 1} \boldsymbol{a}^\top \boldsymbol{x}$. Since the unit ball defined by $\|\boldsymbol{x}\|_{N_1} \leq 1$ is a compact region and the map $\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x}$ is continuous, we know the supremum is attained at a certain point $\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$.

First we consider an vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$ and show that $\boldsymbol{x}^*_{CF} = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ is an optimal counterfactual. We use this $\boldsymbol{v}$ and $d_{\boldsymbol{x}_F}$ to show that the optimality conditions described in Corollary 9 are met. For the first condition we have

$$\begin{aligned}\boldsymbol{a}^\top \boldsymbol{x}^*_{CF} &= \boldsymbol{a}^\top(\boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}) = \boldsymbol{a}^\top \boldsymbol{x}_F + \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F}{\|\boldsymbol{a}\|^*_{N_1}} \boldsymbol{a}^\top \boldsymbol{v} = \boldsymbol{a}^\top \boldsymbol{x}_F + \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F}{\|\boldsymbol{a}\|^*_{N_1}} \|\boldsymbol{a}\|^*_{N_1} \\ &= \boldsymbol{a}^\top \boldsymbol{x}_F + b - \boldsymbol{a}^\top \boldsymbol{x}_F = b.\end{aligned}$$

The second and third conditions require a $\lambda \in \mathbb{R}$. Set $\lambda = \frac{\|\boldsymbol{v}\|_{N_1}}{\boldsymbol{a}^\top \boldsymbol{v}} \operatorname{sgn}(d_{\boldsymbol{x}_F})$. Then, the third condition is met:

$$\|\lambda \boldsymbol{a}\|^*_{N_1} = |\lambda| \|\boldsymbol{a}\|^*_{N_1} = |\lambda| \boldsymbol{a}^\top \boldsymbol{v} = \frac{\|\boldsymbol{v}\|_{N_1}}{|\boldsymbol{a}^\top \boldsymbol{v}|} \boldsymbol{a}^\top \boldsymbol{v} \leq \|\boldsymbol{v}\|_{N_1} \leq 1.$$

Lastly, the second optimality condition is also satisfied:

$$\begin{aligned}\lambda \boldsymbol{a}^\top(\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F) &= \lambda d_{\boldsymbol{x}_F} \boldsymbol{a}^\top \boldsymbol{v} = \frac{\|\boldsymbol{v}\|_{N_1}}{\boldsymbol{a}^\top \boldsymbol{v}} \operatorname{sgn}(d_{\boldsymbol{x}_F}) d_{\boldsymbol{x}_F} \boldsymbol{a}^\top \boldsymbol{v} \\ &= |d_{\boldsymbol{x}_F}| \|\boldsymbol{v}\|_{N_1} = \|d_{\boldsymbol{x}_F} \boldsymbol{v}\|_{N_1} = \|\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F\|_{N_1}.\end{aligned}$$

We conclude that $\boldsymbol{x}^*_{CF} = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ is an optimal counterfactual.

Second, we consider an optimal counterfactual $\boldsymbol{x}^*_{CF} = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ and will show that $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$ with $\|\boldsymbol{v}\|_{N_1} \leq 1$. Since $\boldsymbol{x}^*_{CF}$ is an optimal counterfactual, Corollary 9 states that the two optimality conditions are met. The first condition tells us

$$\boldsymbol{a}^\top \boldsymbol{x}^*_{CF} = \boldsymbol{a}^\top(\boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}) = \boldsymbol{a}^\top \boldsymbol{x}_F + \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F}{\|\boldsymbol{a}\|^*_{N_1}} \boldsymbol{a}^\top \boldsymbol{v} = b.$$

Rearranging the terms yields

$$\frac{\boldsymbol{a}^\top \boldsymbol{v}}{\|\boldsymbol{a}\|^*_{N_1}} = \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F}{b - \boldsymbol{a}^\top \boldsymbol{x}_F} = 1.$$

Hence, we have $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$. The second optimality condition tells us that there exists a $\lambda \in \mathbb{R}$ such that the following relation holds:

$$\lambda \boldsymbol{a}^\top(\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F) = \lambda d_{\boldsymbol{x}_F} \boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{x}^*_{CF} - \boldsymbol{x}_F\|_{N_1} = \|d_{\boldsymbol{x}_F} \boldsymbol{v}\|_{N_1} = |d_{\boldsymbol{x}_F}| \|\boldsymbol{v}\|_{N_1}.$$

Using the fact that $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$, we can rearrange the terms to find

$$\lambda = \operatorname{sgn}(d_{\boldsymbol{x}_F}) \frac{\|\boldsymbol{v}\|_{N_1}}{\|\boldsymbol{a}\|^*_{N_1}}.$$

The last optimality condition implies a bound on $\|\lambda \boldsymbol{a}\|^* \leq 1$:

$$\|\lambda \boldsymbol{a}\|^*_{N_1} = |\lambda| \|\boldsymbol{a}\|^*_{N_1} = |\lambda| \|\boldsymbol{a}\|^*_{N_1} = \frac{\|\boldsymbol{v}\|_{N_1}}{\|\boldsymbol{a}\|^*_{N_1}} \|\boldsymbol{a}\|^*_{N_1} = \|\boldsymbol{v}\|_{N_1} \leq 1.$$

Therefore, it must hold that $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|^*_{N_1}$ □

Lemma 13. Note that Lemma 11 proves the existence of $\boldsymbol{v}$. Since the set $\{\boldsymbol{v}, \boldsymbol{v}^2, \ldots, \boldsymbol{v}^p\}$ is linearly independent, we know that for $i = 2, \ldots, p$ the counterfactuals are nonzero, *i.e.*, $\boldsymbol{v}^{\boldsymbol{j}}_{CF} = \boldsymbol{v}^j + d_{\boldsymbol{v}^j}\boldsymbol{v} \neq \boldsymbol{0}$.

(i) Assume $\boldsymbol{v}_{CF} \neq \boldsymbol{0}$. Consider $\lambda, \lambda_2, \ldots, \lambda_p$ such that

$$\lambda \boldsymbol{v}_{CF} + \sum_{i=2}^{p} \lambda_i \boldsymbol{v}^i_{CF} = \boldsymbol{0}.$$

By construction of the counterfactuals, we have

$$\lambda \boldsymbol{v}_{CF} + \sum_{i=2}^{p} \lambda_i \boldsymbol{v}_{CF}^i = \lambda \left(\boldsymbol{v} + d_{\boldsymbol{v}} \boldsymbol{v}\right) + \sum_{i=2}^{p} \lambda_i \left(\boldsymbol{v}^i + d_{\boldsymbol{v}^i} \boldsymbol{v}\right)$$
$$= \boldsymbol{v} \left(\lambda + \lambda d_{\boldsymbol{v}} + \sum_{i=2}^{p} \lambda_i d_{\boldsymbol{v}^i}\right) + \sum_{i=2}^{p} \lambda_i \boldsymbol{v}^i = \boldsymbol{0}.$$

This implies that $\lambda_2 = \ldots = \lambda_p = 0$, since otherwise $\boldsymbol{v}$ is a linear combination of vectors $\boldsymbol{v}^2, \ldots, \boldsymbol{v}^p$ which contradicts with V being a basis. The resulting equation is $\lambda(1 + d_{\boldsymbol{v}})\boldsymbol{v} = \lambda \boldsymbol{v}_{CF} = \boldsymbol{0}$. Since $\boldsymbol{v}_{CF} \neq \boldsymbol{0}$, it must hold that $\lambda = 0$. Hence, the set $\{\boldsymbol{v}_{CF}, \boldsymbol{v}_{CF}^2, \ldots, \boldsymbol{v}_{CF}^p\}$ is linearly independent.

(ii) Assume $\boldsymbol{v}_{CF} = \boldsymbol{0}$ and consider $\lambda_2, \ldots, \lambda_p$ such that

$$\sum_{i=2}^{p} \lambda_i \boldsymbol{v}_{CF}^i = \boldsymbol{0}.$$

By construction of the counterfactuals, we have

$$\sum_{i=2}^{p} \lambda_i \boldsymbol{v}_{CF}^i = \sum_{i=2}^{p} \lambda_i \left(\boldsymbol{v}^i + d_{\boldsymbol{v}^i} \boldsymbol{v}\right) = \sum_{i=2}^{p} \lambda_i \boldsymbol{v}^i + \boldsymbol{v} \sum_{i=2}^{p} \lambda_i d_{\boldsymbol{v}^i} = \boldsymbol{0}.$$

This implies that $\lambda_2 = \ldots = \lambda_p = 0$, since otherwise $\boldsymbol{v}$ is a linear combination of vectors $\boldsymbol{v}^2, \ldots, \boldsymbol{v}^p$ which contradicts with V being a basis. Hence, the set $\{\boldsymbol{v}_{CF}^2, \ldots, \boldsymbol{v}_{CF}^p\}$ is linearly independent.

□

Theorem 14. It follows directly from Lemmas 11 and 13 that Algorithm 1 finds p linearly independent vectors on the hyperplane to retrieve the original hyperplane exactly. Since we assume $\lambda \boldsymbol{e}^1$ is not on the hyperplane, it takes one counterfactual query to find the direction of the counterfactuals, $\boldsymbol{v}$. Then only p counterfactual queries are needed for the newly created basis V to obtain a linearly independent system of equations to retrieve the original hyperplane exactly. □

A.3 Robust Counterfactuals

Lemma 16. Without loss of generality, we let $\boldsymbol{x}_F$ be classified as *'No'* and reformulate (2) as

$$\min_{\boldsymbol{x}_{RCF}} \ \|\boldsymbol{x}_{RCF} - \boldsymbol{x}\|_{N_1}$$
$$s.t. \quad \boldsymbol{a}^\top (\boldsymbol{x}_{RCF} + \boldsymbol{s}) \geq b \quad \forall \boldsymbol{s} \in \mathcal{S}.$$

Using the classical robust optimization reformulation, we can rewrite the infinite set of constraints as

$$\boldsymbol{a}^\top (\boldsymbol{x}_{RCF} + \boldsymbol{s}) \geq b \quad \forall \boldsymbol{s} \in \mathcal{S} \qquad \Leftrightarrow \qquad \boldsymbol{a}^\top \boldsymbol{x}_{RCF} + \min_{\boldsymbol{s}: \|\boldsymbol{s}\|_{N_2} \leq \rho} \boldsymbol{a}^\top \boldsymbol{s} \geq b$$
$$\Leftrightarrow \qquad \boldsymbol{a}^\top \boldsymbol{x}_{RCF} - \rho \|\boldsymbol{a}\|_{N_2}^* \geq b,$$

where $\|\cdot\|_{N_2}^*$ is the dual norm of $\|\cdot\|_{N_2}$. The latter constraint is linear in the optimization variable $\boldsymbol{x}_{RCF}$. Since the objective function is convex, we know that every optimal solution $\boldsymbol{x}^*$ must fulfill the KKT-conditions

$$b - \boldsymbol{a}^\top \boldsymbol{x}^* + \rho \|\boldsymbol{a}\|_{N_2}^* \leq 0,$$
$$\boldsymbol{0} \in \partial_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*),$$
$$\lambda^* (b - \boldsymbol{a}^\top \boldsymbol{x}^* + \rho \|\boldsymbol{a}\|_{N_2}^*) = 0,$$
$$\lambda^* \geq 0,$$

where λ^* is a dual optimal solution and $\mathcal{L}$ is the Lagrangian dual function

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \|\boldsymbol{x} - \boldsymbol{x}_F\|_{N_1} + \lambda(b - \boldsymbol{a}^\top \boldsymbol{x} + \rho\|\boldsymbol{a}\|_{N_2}^*).$$

We have $\partial_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \lambda) = \partial f(\boldsymbol{x} - \boldsymbol{x}_F) - \lambda\boldsymbol{a}$. Note that since $\boldsymbol{x}_F \neq \boldsymbol{x}_{RCF}^*$, we have $\boldsymbol{0} \notin \partial f(\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F)$ and hence, $\lambda^* > 0$ must hold. Applying this to the third KKT results in

$$b - \boldsymbol{a}^\top \boldsymbol{x}^* + \rho\|\boldsymbol{a}\|_{N_2}^* = 0,$$

which makes the first condition redundant. This proves the result. □

Theorem 17. First, we query $\boldsymbol{x}_F$ such that we know the value of $q_F(\boldsymbol{x}_F)$. Then, we apply Lemma 16 to the case where the norm is differentiable. Then, there exists an $\lambda^* > 0$ such that

$$\begin{aligned}\lambda^*\boldsymbol{a} &= \nabla f(\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F),\\ b &= \boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* + q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}.\end{aligned}$$

By defining $\hat{\boldsymbol{a}} = \nabla f(\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F)$ and $\hat{b} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_{RCF}^* + q_F(\boldsymbol{x}_F)\rho\|\hat{\boldsymbol{a}}\|_{N_2}$ we have $\lambda^*\boldsymbol{a} = \hat{\boldsymbol{a}}$ and $\lambda^* b = \lambda^*\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* + \lambda^* q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2} = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_{RCF}^* + q_F(\boldsymbol{x}_F)\rho\|\hat{\boldsymbol{a}}\|_{N_2} = \hat{b}$. Hence, we $\frac{1}{\lambda^*}(\hat{\boldsymbol{a}}, \hat{b}) = (\boldsymbol{a}, b)$ for a $\lambda^* \neq 0$ and the hyperplane given by $\boldsymbol{a}, b$ is equivalent to the original hyperplane with parameters $\boldsymbol{a}, b$ which proves the result. □

Lemma 20. We note that $\boldsymbol{a} \neq \boldsymbol{0}$, hence $\|\boldsymbol{a}\|_{N_1}^* \neq \boldsymbol{0}$ and $d_{\boldsymbol{x}_F}$ is well defined. Besides, because $q_F(\boldsymbol{x}_F) \neq 0$ we have $d_{\boldsymbol{x}_F} > 0$. Moreover, we have $\|\boldsymbol{a}\|_{N_1}^* := \sup_{\|\boldsymbol{x}\|_{N_1} \leq 1} \boldsymbol{a}^\top \boldsymbol{x}$. Since the unit ball defined by $\|\boldsymbol{x}\|_{N_1} \leq 1$ is a compact space and the map $\boldsymbol{x} \mapsto \boldsymbol{a}^\top \boldsymbol{x}$ is continuous, we know the supremum is attained at a certain point $\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$.

First, we consider a vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$ and show that $\boldsymbol{x}_{RCF}^* = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ is an optimal robust counterfactual. We show the optimality conditions described in Corollary 9 are met. For the first condition we have

$$\begin{aligned}\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* &= \boldsymbol{a}^\top(\boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}) = \boldsymbol{a}^\top \boldsymbol{x}_F + \left(\frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*}{\|\boldsymbol{a}\|_{N_1}^*}\right)\boldsymbol{a}^\top \boldsymbol{v}\\ &= \boldsymbol{a}^\top \boldsymbol{x}_F + b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*\\ &= b - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*.\end{aligned}$$

The second and third conditions require a $\lambda \in \mathbb{R}$. Set $\lambda = \frac{\|\boldsymbol{v}\|_{N_1}}{\boldsymbol{a}^\top \boldsymbol{v}}\operatorname{sgn}(d_{\boldsymbol{x}_F})$. Then, the third condition is met:

$$\|\lambda\boldsymbol{a}\|_{N_1}^* = |\lambda|\|\boldsymbol{a}\|_{N_1}^* = |\lambda|\boldsymbol{a}^\top \boldsymbol{v} = \frac{\|\boldsymbol{v}\|_{N_1}}{|\boldsymbol{a}^\top \boldsymbol{v}|}\boldsymbol{a}^\top \boldsymbol{v} \leq \|\boldsymbol{v}\|_{N_1} \leq 1.$$

Lastly, the second optimality condition is also met:

$$\begin{aligned}\lambda\boldsymbol{a}^\top(\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F) &= \lambda d_{\boldsymbol{x}_F}\boldsymbol{a}^\top \boldsymbol{v} = \frac{\|\boldsymbol{v}\|_{N_1}}{\boldsymbol{a}^\top \boldsymbol{v}}\operatorname{sgn}(d_{\boldsymbol{x}_F})d_{\boldsymbol{x}_F}\boldsymbol{a}^\top \boldsymbol{v}\\ &= |d_{\boldsymbol{x}_F}|\|\boldsymbol{v}\|_{N_1} = \|d_{\boldsymbol{x}_F}\boldsymbol{v}\|_{N_1} = \|\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F\|_{N_1}.\end{aligned}$$

We conclude that $\boldsymbol{x}_{RCF}^* = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ is an optimal robust counterfactual.

Second, we consider an optimal counterfactual $\boldsymbol{x}_{RCF}^* = \boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}$ and show that $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$ with $\|\boldsymbol{v}\|_{N_1} \leq 1$. Since $\boldsymbol{x}_{RCF}^*$ is an optimal counterfactual, Corollary 18 implies the corresponding optimality conditions are met. The first condition tells us

$$\begin{aligned}\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* &= \boldsymbol{a}^\top(\boldsymbol{x}_F + d_{\boldsymbol{x}_F}\boldsymbol{v}) = \boldsymbol{a}^\top \boldsymbol{x}_F + \left(\frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*}{\|\boldsymbol{a}\|_{N_1}^*}\right)\boldsymbol{a}^\top \boldsymbol{v}\\ &= b - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*.\end{aligned}$$

Rearranging the terms yields

$$\frac{\boldsymbol{a}^\top \boldsymbol{v}}{\|\boldsymbol{a}\|_{N_1}^*} = \frac{b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*}{b - \boldsymbol{a}^\top \boldsymbol{x}_F - q_F(\boldsymbol{x}_F)\rho\|\boldsymbol{a}\|_{N_2}^*} = 1.$$

Hence, we have $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$. The second optimality condition tells us there exists a $\lambda \in \mathbb{R}$ such that the following relation holds.

$$\lambda \boldsymbol{a}^\top (\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F) = \lambda d_{\boldsymbol{x}_F} \boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{x}_{RCF}^* - \boldsymbol{x}_F\|_{N_1} = \|d_{\boldsymbol{x}_F} \boldsymbol{v}\|_{N_1} = |d_{\boldsymbol{x}_F}| \|\boldsymbol{v}\|_{N_1}$$

Using the fact that $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$, we can rearrange the terms to find

$$\lambda = \operatorname{sgn}(d_{\boldsymbol{x}_F}) \frac{\|\boldsymbol{v}\|_{N_1}}{\|\boldsymbol{a}\|_{N_1}^*}.$$

The last optimality condition implies a bound on $\|\lambda \boldsymbol{a}\|^* \leq 1$.

$$\|\lambda \boldsymbol{a}\|_{N_1}^* = |\lambda| \|\boldsymbol{a}\|_{N_1}^* = |\lambda| \|\boldsymbol{a}\|_{N_1}^* = \frac{\|\boldsymbol{v}\|_{N_1}}{\|\boldsymbol{a}\|_{N_1}^*} \|\boldsymbol{a}\|_{N_1}^* = \|\boldsymbol{v}\|_{N_1} \leq 1.$$

Therefore, it must hold that $\|\boldsymbol{v}\|_{N_1} \leq 1$ and $\boldsymbol{a}^\top \boldsymbol{v} = \|\boldsymbol{a}\|_{N_1}^*$. □

Lemma 23. Without loss of generality, assume that $\boldsymbol{x}_F$ is classified as *'No'*. From Corollary 18 it follows that there exists a $\lambda \in \mathbb{R} \setminus \{0\}$ such that

$$\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* = b + \rho \|\boldsymbol{a}\|_{N_2}^*, \tag{18}$$

$$\lambda \boldsymbol{a}^\top \boldsymbol{v} = 1, \tag{19}$$

$$\|\lambda \boldsymbol{a}\|_{N_1}^* \leq 1. \tag{20}$$

Furthermore, from the second condition the definition of $\boldsymbol{v}$ and since $\boldsymbol{a}^\top \boldsymbol{x}_{RCF}^* \geq b \geq \boldsymbol{a}^\top \boldsymbol{x}_F$, it follows that $\lambda > 0$ must hold. From the latter conditions and the equivalence of the norms, we obtain for the inner product of $\boldsymbol{a}$ and $\bar{\boldsymbol{x}}$

$$\begin{aligned} \boldsymbol{a}^\top \bar{\boldsymbol{x}} &= \boldsymbol{a}^\top (\boldsymbol{x}_{RCF}^* - d\boldsymbol{v}) \\ &= b + \rho \|\boldsymbol{a}\|_{N_2}^* - d\lambda^{-1} \\ &\leq b + \rho C \|\boldsymbol{a}\|_{N_1}^* - d\lambda^{-1} \\ &= b + \lambda^{-1} (\rho C - d), \end{aligned} \tag{21}$$

where the second equality follows from (18) and (19), the first inequality follows from the equivalence of the dual norms, and the last equality follows from (19) and (20) since $\lambda^{-1} = \boldsymbol{a}^\top \boldsymbol{v} \leq \|\boldsymbol{a}\|_{N_1}^* \leq \lambda^{-1}$. Hence for $d \geq \rho C$ we have $\boldsymbol{a}^\top \bar{\boldsymbol{x}} \leq b$. □

Corollary 24. The statements (i) and (ii) follow from applying Lemma 23 together with classical results for dual ℓ_p-norms and equivalence between ℓ_p-norms. Statement (iii) follows since the inequality in (21) is an equality with $C = 1$, if $N_1 = N_2$. □