Online Supplement to Finding Regions of Counterfactual Explanations via Robust Optimization

Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, Ş. İlker Birbil, Dick den Hertog
Amsterdam Business School, University of Amsterdam, 1018TV Amsterdam, Netherlands
d.maragno@uva.nl j.kurtz@uva.nl t.e.rober@uva.nl r.goedhart2@uva.nl s.i.birbil@uva.nl d.denhertog@uva.nl

Appendix A: How to choose the robustness budget ρ

When determining the robustness budget ρ, it is essential to weigh the balance between the distance of the resulting CE to the factual instance and its robustness against small perturbations. When the underlying distribution of the CE perturbations is known, we can select ρ such that a certain probabilistic guarantee is satisfied. However, in many real-world situations, we do not have access to distributional information regarding the perturbations. In this case, decision-makers can derive a Pareto front based on the CE's proximity to the factual instance and its robustness against changes.

A.1. Probability guarantee

Probability guarantees regarding the invalidation of recourse depend on both the specific predictive model and the type of uncertainty set in use. For a comprehensive overview of robust reformulations with probabilistic guarantees in the context of linear models, we refer to (Ben-Tal et al. 2009). In this section, we present a more generic formulation that is applicable to both linear and nonlinear models. However, it should be noted that it is a more conservative approach since it only considers the probability that the realized recourse will be in the uncertainty set while (large) part(s) of the uncertainty set may not be near the decision boundary, depending on the model structure.

If we assume that the realized/actual perturbation $\hat{\boldsymbol{s}}$ is distributed according to $\hat{\boldsymbol{s}} \sim N(0, \sigma^2 I)$, where I is the identity matrix, then we can approach the probabilistic guarantee from three angles:

1. What is the probability (α) that the realized recourse is in the defined uncertainty set $\mathcal{S} = \{\boldsymbol{s} \in \mathbb{R}^n \mid ||\boldsymbol{s}|| \leq \rho\}$, given ρ and σ?
2. What value of ρ is required to have α probability that the realized recourse is in the uncertainty set $\mathcal{S} = \{\boldsymbol{s} \in \mathbb{R}^n \mid ||\boldsymbol{s}|| \leq \rho\}$, given α and σ?
3. How large can the standard deviation of the perturbations (σ) be to have α probability of the realized recourse being in the uncertainty set $\mathcal{S} = \{\boldsymbol{s} \in \mathbb{R}^n \mid ||\boldsymbol{s}|| \leq \rho\}$, given α and ρ?

These answers depend on the chosen norm for the uncertainty set. Note that a prediction region with probability α for $\hat{\boldsymbol{s}}$ is given by $\tilde{\mathcal{S}} = \{\boldsymbol{s} \in \mathbb{R}^n \mid \boldsymbol{s}^T\boldsymbol{s} \leq \sigma^2\chi^2_k(\alpha)\}$ (Chew 1966), such that for the l_2 norm the questions above boil down to

$$\alpha = F_{\chi^2_k}\left(\frac{\rho^2}{\sigma^2}\right), \tag{68}$$

$$\rho = \sigma\sqrt{\chi^2_k(\alpha)}, \tag{69}$$

$$\sigma = \rho/\sqrt{\chi^2_k(\alpha)}, \tag{70}$$

where $\chi^2_k(\alpha)$ and is the quantile-function for the probability α for the chi-square distribution with k degrees of freedom, k is the dimension (number of features) of $\hat{\boldsymbol{s}}$, and $F_{\chi^2_k}$ is the cumulative distribution function for the chi-square distribution with k degrees of freedom.

For the l_∞-norm, first note that $\hat{\boldsymbol{s}}$ is a vector of k i.i.d. $N(0, \sigma^2)$ variables. The probability that the realized recourse is in the defined uncertainty set for this norm is then equal to the probability that all k variables are between $-\rho$ and ρ, i.e.: $P(\hat{\boldsymbol{s}} \in \mathcal{S}) = [\Phi(\rho/\sigma) - \Phi(-\rho/\sigma)]^k = [2\Phi(\rho/\sigma) - 1]^k$, where $\Phi(x)$ represents the standard normal cumulative distribution function. Equating $P(\hat{\boldsymbol{s}} \in \mathcal{S})$ to α then yields:

$$\alpha = \left[2\Phi\left(\frac{\rho}{\sigma}\right) - 1\right]^k, \tag{71}$$

$$\rho = \sigma\Phi^{-1}\left(\frac{\alpha^{1/k} + 1}{2}\right), \tag{72}$$

$$\sigma = \rho/\Phi^{-1}\left(\frac{\alpha^{1/k} + 1}{2}\right), \tag{73}$$

where $\Phi^{-1}(x)$ represents the inverse of the standard normal cumulative distribution function.

A.2. Pareto front

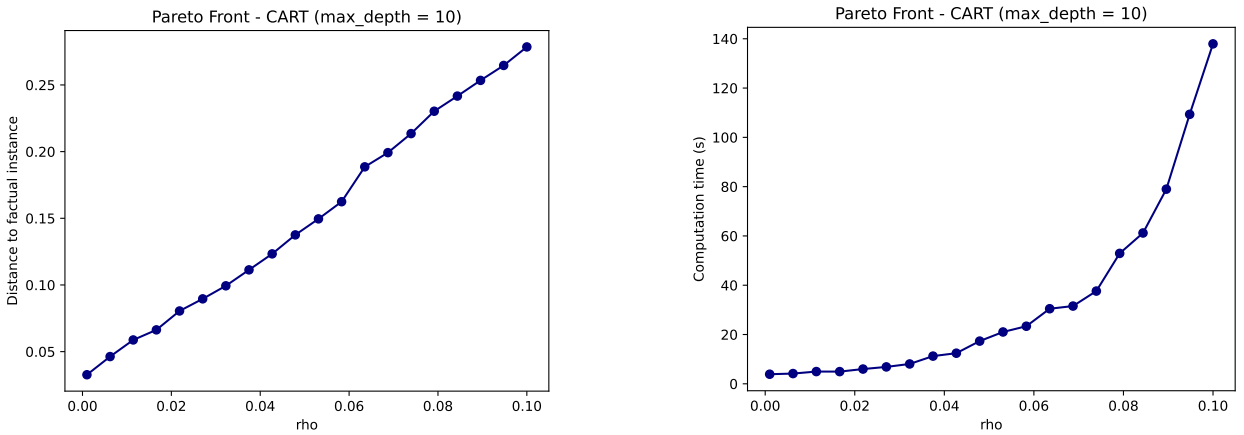
The Pareto front represents a set of CE solutions that cannot be improved in one objective without degrading performance in the other. In our context:

- Objective 1: Proximity of the CE to the Factual Instance – This measures the ℓ_1-norm distance between the CE and the factual instance.
- Objective 2: Robustness (ρ) - This quantifies the robustness of the CE against perturbations.

In Figure 1a, we report the Pareto front for a decision tree with a max depth of 10 and ℓ_∞-norm uncertainty set. The outcomes are averaged across 20 different cases. When we increase the value of ρ, the distance to the factual instance also increases, occasionally experiencing jumps, see for example the jump in distance for values of ρ around 0.06, which can be attributed to changes in the leaf nodes. Also, the computation time increases with ρ given the increasing difficulty in finding a region that is large enough to contain the ρ-robust CE, see Figure 1b.

Appendix B: Linear Models

Although Algorithm 1 could still be used for the linear models, there is a well-known easier, and more efficient dual approach to solve model (3)-(4). We review this approach briefly here for the completeness of our discussion.



(a) (b)

Figure 1 **Pareto fronts obtained comparing (a) distance to the factual instance (y-axis) and ρ (x-axis) and (b) computation time required to find a robust counterfactual (y-axis) and ρ (x-axis). The results reported are averaged over 20 factual instances and obtained using a decision tree with a max-depth of 10 trained on the diabetes dataset.**

In the case of linear models, such as logistic regression (LR) or linear support vector machines (SVM), the validity constraint (4) can be formulated as

$$\boldsymbol{\beta}^\top(\boldsymbol{x}+\boldsymbol{s})+\beta_0 \geq \tau, \quad \forall \boldsymbol{s} \in \mathcal{S}, \tag{74}$$

where $\boldsymbol{\beta} \in \mathbb{R}^n$ is the coefficient vector and $\beta_0 \in \mathbb{R}$ is the intercept. Then, these constraints can be equivalently reformulated as

$$\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_0 + \min_{\boldsymbol{s}\in\mathcal{S}} \beta^\top \boldsymbol{s} \geq \tau.$$

Since $\mathcal{S}$ is given in the form (5), the latter is equivalent to

$$\boldsymbol{\beta}^\top \boldsymbol{x} - \rho||\boldsymbol{\beta}||^* + \beta_0 \geq \tau, \tag{75}$$

where $\|\cdot\|^*$ is the dual norm of the norm used in the definition of $\mathcal{S}$. The constant term $\rho||\boldsymbol{\beta}||^*$ ensures that the constraint (74) holds for all $\boldsymbol{s} \in \mathcal{S}$. Note that constraint (75) remains linear in $\boldsymbol{x}$ independently of $\mathcal{S}$. For more details see, *e.g.*, Dominguez-Olmedo et al. (2021), Bertsimas et al. (2019), Xu et al. (2008).

Appendix C: Proof of Lipschitz Continuity of Neural Networks

Suppose that we have a trained ℓ-layer neural network constructed with ReLU activation functions. If we denote the resulting functional by $f : \mathbb{R}^{n_0} \mapsto \mathbb{R}^{n_\ell}$, then we can write

$$f(\boldsymbol{x}) = \sigma_\ell(W_\ell \sigma_{\ell-1}(W_{\ell-1}\sigma_{\ell-2}\dots W_2\sigma_1(W_1\boldsymbol{x})\dots)), \tag{76}$$

where $\sigma_m : \mathbb{R}^{n_m} \mapsto \mathbb{R}^{n_m}$, $m = 1,\dots,\ell$ stands for the vectorized ReLU functions, and $W_m \in \mathbb{R}^{n_m} \times \mathbb{R}^{n_{m-1}}$, for $m = 1,\dots,\ell$, are the weight matrices. This shows that f is simply the composition of linear and component- as well as piece-wise linear functions.

Given any two Lipschitz continuous functions $g : \mathbb{R}^p \mapsto \mathbb{R}^q$ and $h : \mathbb{R}^q \mapsto \mathbb{R}^s$ with respective Lipschitz constants L_g and L_h, the composition $h \circ g : \mathbb{R}^p \mapsto \mathbb{R}^p$ is Lipschitz continuous with Lipschitz constant $L_h L_g$. This simply follows from observing for a pair of vectors $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$ that

$$\|h \circ g(\boldsymbol{y}) - h \circ g(\boldsymbol{z})\| \leq L_h \|g(\boldsymbol{y}) - g(\boldsymbol{z})\| \leq L_h L_g \|\boldsymbol{y} - \boldsymbol{z}\|. \tag{77}$$

Next, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{n_m}$, we have

$$\|W_m \boldsymbol{u} - W_m \boldsymbol{v}\| \leq \|W_m\|_s \|\boldsymbol{u} - \boldsymbol{v}\|. \tag{78}$$

where $\|\cdot\|_s$ is the spectral norm. As the Lipschitz constant for any vectorized ReLU function is one, we obtain the desired result by

$$\|f(\bar{\boldsymbol{x}}) - f(\tilde{\boldsymbol{x}})\| \leq \prod_{m=1}^{\ell} \|W_m\|_s \|\bar{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| \tag{79}$$

for any pair $\bar{\boldsymbol{x}}, \tilde{\boldsymbol{x}} \in \mathbb{R}^{n_0}$. □

Appendix D: Performances of Predictive Models

	LR	**CART**			**RF**					**GBM**					**NN**			
		3	**5**	**10**	**5**	**10**	**20**	**50**	**100**	**5**	**10**	**20**	**50**	**100**	**(10)**	**(10, 10, 10)**	**(50)**	**(100)**
BANKNOTE																		
Train	0.97	0.94	0.98	1.00	0.96	0.97	0.96	0.97	0.97	0.99	0.99	1.00	1.00	1.00	0.98	0.99	1.00	1.00
Test	0.96	0.89	1.00	1.00	0.96	0.93	0.93	0.96	0.93	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIABETES																		
Train	0.77	0.77	0.84	0.97	0.77	0.77	0.79	0.80	0.80	0.80	0.83	0.87	0.93	0.99	0.78	0.79	0.80	0.82
Test	0.72	0.77	0.72	0.69	0.72	0.69	0.69	0.67	0.71	0.59	0.67	0.69	0.59	0.62	0.69	0.64	0.73	0.77
IONOSPHERE																		
Train	0.88	0.93	0.97	1.00	0.93	0.95	0.94	0.96	0.96	0.99	1.00	1.00	1.00	1.00	0.98	0.99	0.99	0.99
Test	0.83	0.88	0.83	0.83	0.88	0.92	0.92	0.92	0.92	0.88	0.88	0.92	0.92	0.92	0.89	0.92	0.88	0.88

Table 1 Train and test accuracy scores of the predictive models used for the experiments.

Table 1 displays the accuracy score of each predictive model employed in the experiments, with their performance being reported for both the training and testing sets.

References

Ben-Tal A, Ghaoui L, Nemirovski A (2009) *Robust Optimization*. ISBN 9781400831050, URL `http://dx.doi.org/10.1515/9781400831050`.

Bertsimas D, Dunn J, Pawlowski C, Zhuo YD (2019) Robust classification. *INFORMS Journal on Optimization* 1(1):2–34.

Chew V (1966) Confidence, prediction, and tolerance regions for the multivariate normal distribution. *Journal of the American Statistical Association* 61(315):605–617.

Dominguez-Olmedo R, Karimi AH, Schölkopf B (2021) On the adversarial robustness of causal algorithmic recourse. URL `http://dx.doi.org/10.48550/ARXIV.2112.11313`.

Xu H, Caramanis C, Mannor S (2008) Robust regression and lasso. *Advances in Neural Information Processing Systems* 21.