GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations

Dan Ley [*] [1]   Saumitra Mishra [2]   Daniele Magazzeni [2]

MI2 Seminar

Piotr Wilczyński

May 6th, 2024

Authors

Dan Ley
Harvard University

Saumitra Mishra
J.P. Morgan AI Research

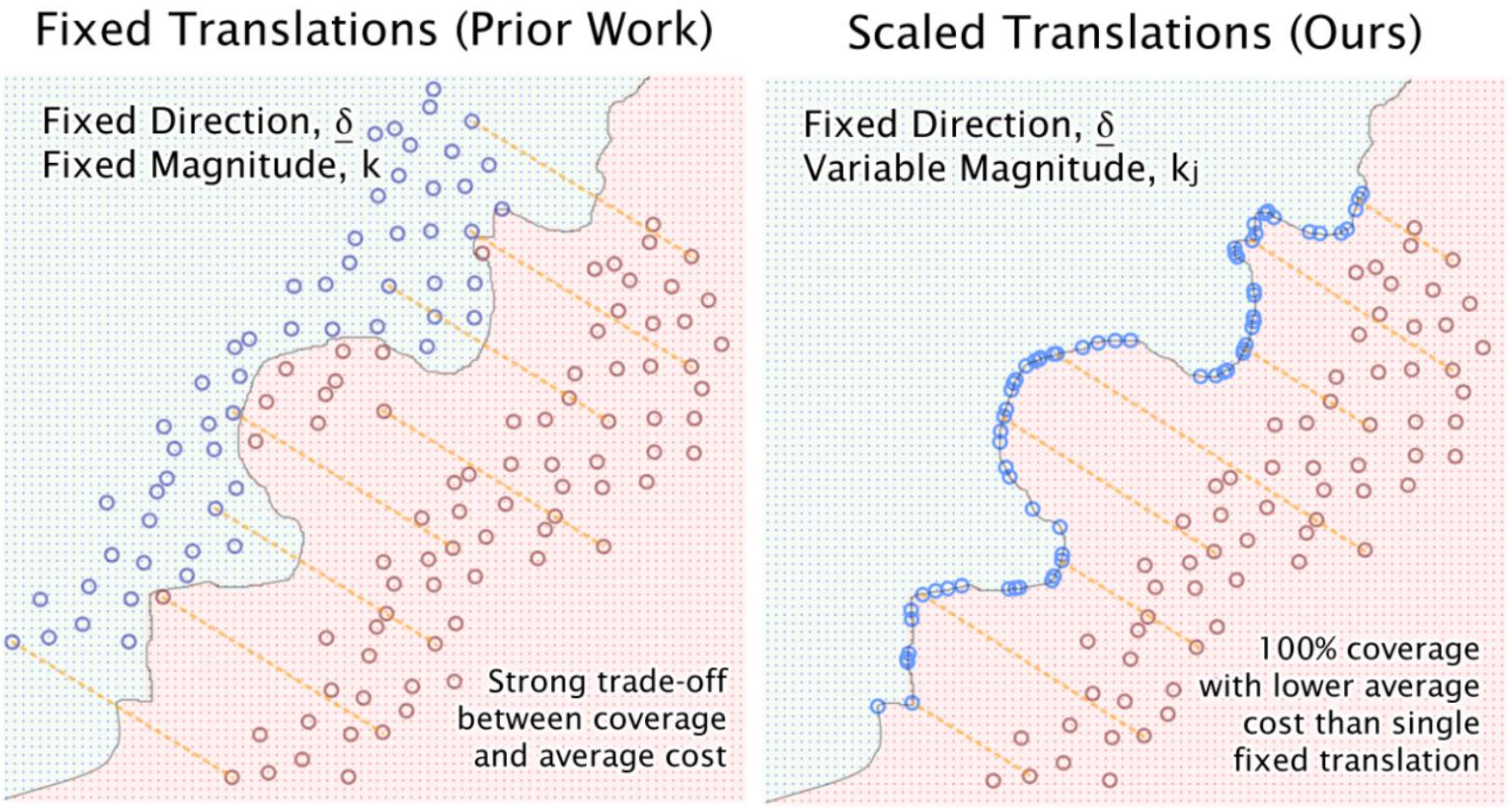
Daniele Magazzeni
J.P. Morgan AI Research

International Conference on Machine Learning 2023

Motivation

- Inability of CEs to provide explanations beyond the local or instance-level
- A local CE for a specific sample cannot represent the bias of the entire model
- Only few works provide global explanation frameworks that are both reliable and computationally tractable
- Practitioners are requesting more efficient and interactive explainability tools
- It is not evident that aggregating local explanations would scale well or lead to reliable conclusions about a model's behaviour
- In prior work, GCEs simply took the same form as CEs, but applied to an entire group of inputs – such formulation fails to overcome the trade-off between coverage and cost
- Relaxed objective, where each GCE represents just the translation direction, successfully overcomes this limitation

Solution Intuition



Definitions

- counterfactuals – the altered inputs
- counterfactual explanations (CE) – any representation of the change required
- global counterfactual explanation (GCE) – global direction along which a group of inputs may travel to alter their predictions (translation direction)
- δ– fixed translation direction
- k_j–variable magnitude

Contribution

I. Proposing a framework that permits GCEs to have variable magnitudes while preserving a fixed translation direction mitigating the trade-off between coverage and cost

II. Proving that arbitrary translations on one-hot encodings (categorical data) can be expressed using If/Then rules

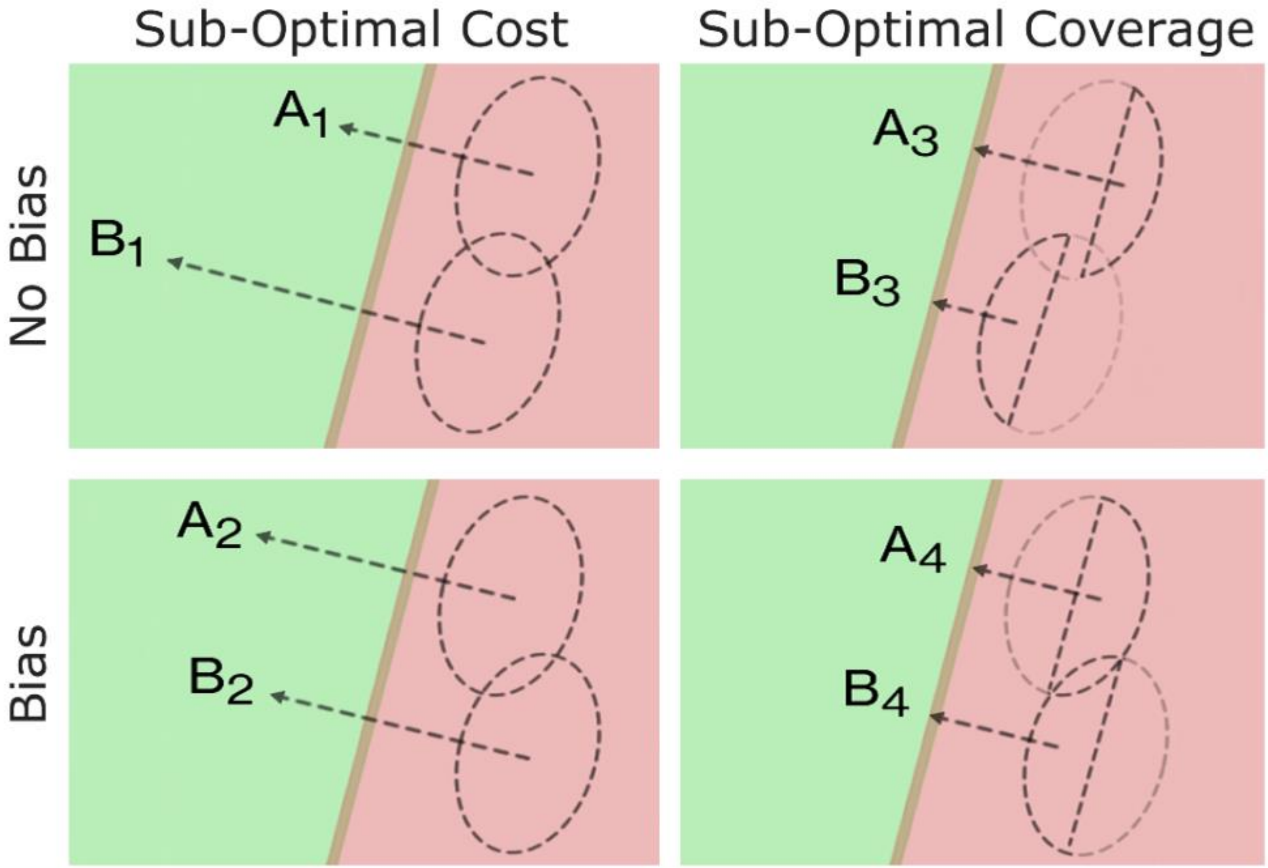
III. Demonstrating that GLOBE-CE outperforms competing methods in coverage, cost, and runtime

Reliability vs Efficiency

- **Reliable** GCEs are those that can be used to draw accurate conclusions of a model’s behaviour (maximum coverage and minimum average costs).

- **Efficiency** is defined in relation to the average CPU time taken in computing GCEs.

Reliability



Representation: Scaled Translation Vectors

- For inputs that belong to a particular subgroup $\underline{x} \in \mathcal{X}$, we can apply a translation $\underline{\delta}$ with scalar k such that $\underline{x}_{CF} = \underline{x} + k\underline{\delta}$ is a valid counterfactual
- For each $\underline{x} \in \mathcal{X}$, framework computes the respective minimum value of k required for recourse
- This approach guarantees improvement with respect to the interpretability to performance trade-off that other methods suffer from

Translations on Categorical Features

- **Goal.** Show that arbitrary translations on one-hot encodings (categorical data) can be expressed using If/Then rules
- **Theorem 4.1.** *Regardless of the feature value of the input, any translation vector that is added to a one-hot categorical input can alternatively be expressed using If/Then rules, with just one unique Then condition.*
- **Theorem 4.2.** *Regardless of the feature value of the input, any translation vector that is scaled by $k \geq 0$ and added to a one-hot categorical input can alternatively be expressed with the first m rules of a sequence.*

Theorem 4.1

- n – number of feature labels
- $\underline{f} = [f_1, f_2, \dots, f_3] \in \{0,1\}^n$ where $\left|\underline{f}\right|_1 = 1$ – one-hot encoded feature vector
- $F = argmax_i(f_i)$
- $\underline{\delta} = [\delta_1, \delta_2, \dots, \delta_3] \in R^n$ – translation vector
- $\Delta = argmax_i(\delta_i)$
- $\underline{g} = \underline{f} + \underline{\delta}$ – post-translation vector
- $G = argmax_i(g_i)$ – final feature value

Note: $g_{i \neq F} = \delta_i$ and $g_F = \delta_F + 1$

Theorem 4.1

- $g_G = max_i(g_i) = max(\delta_F + 1, max_{i \neq F}(\delta_i))$
- For $1 \leq F \leq n$, we now prove that if $G \neq F$ (i.e. a change in feature value occurs), we have the rule “If F, Then Δ”
- Case $F = \Delta$. $g_G = max\left(\delta_\Delta + 1, max_i(\delta_{i \neq \Delta})\right) = \delta_\Delta + 1$
  Hence, $\mathrm{G} = \Delta$ (no rule)
- Case $F \neq \Delta$. $g_G = max(\delta_F + 1, \delta_\Delta)$
  - If $\delta_F + 1 > \delta_{i \neq \Delta}$ then $g_G = \delta_F + 1$ and $\mathrm{G} = F$ (no rule)
  - If $\delta_F + 1 < \delta_{i \neq \Delta}$ then $g_G = \delta_\Delta$ and $\mathrm{G} = \Delta$ (rule “If F, Then Δ”) ■

Theorem 4.2

- k – scalar
- For $i \neq \Delta$ and $k > 0$ **Theorem 4.1** gives that $k\delta_i + 1 < k\delta_\Delta$ yields the rule "If i, Then Δ"
- Thus, if the lower bound $k > \frac{1}{\delta_\Delta - \delta_i}$ is satisfied then $k\underline{\delta}$ induces such a rule
- Let's consider the vector of lower bounds $\underline{k} = [k_1, k_2, \dots, k_n] \in R_+^n$ where $k_{i \neq \Delta} = \frac{1}{\delta_\Delta - \delta_i}$ and $k_\Delta = \infty$
- …

Lemma 4.2.1

- $k_i \leq k_m$ for any $i, m < n$ with $\delta_i \leq \delta_m$
- Lower bounds for i and m are both satisfied if $k > k_m$
- Thus, scaling $\underline{\delta}$ by $k > k_m$ induces the rule corresponding to each feature value i with $\delta_i \leq \delta_m$ ■

Theorem 4.2

- For $k = 0$, we have no rules ($k\underline{\delta} = \underline{0}$)
- Δ_i – index of the i^{th} smallest value in $\underline{\delta}$
- Thus, by **Lemma 4.2.1**, for $m < n$, we have that scaling $\underline{\delta}$ by $k_{\Delta_m} < \mathrm{k} \leq k_{\Delta_{m+1}}$ induces rules for the first m feature values $\Delta_{1 \leq i \leq m}$ ■

GLOBE-CE algorithm

- The major contribution of the GLOBE-CE framework lies in the notion of scaling the magnitudes of translations
- One can interpret a range of magnitudes, though cannot interpret a range of directions so easily

GLOBE-CE algorithm

- Explanations are learned by adopting methods from instance-level CE, generalising for any CE algorithm $G(B, \mathcal{X}, n)$ that considers, at a minimum, the model B being explained, the inputs requiring explanations $\mathcal{X}$, and the number n of returned GCEs $\underline{\delta}_1, \underline{\delta}_2, \dots, \underline{\delta}_n = \Delta$
- GLOBE-CE scales the i^{th} GCE $\underline{\delta}_i$ over a range of m scalars $\underline{k} = k_1, k_2, \dots, k_m$, repeating over all $1 \leq i \leq n$ GCEs and returning the counterfactuals $\mathcal{X}'$, the predictions $Y' \in \{0,1\}^{n \times m \times |\mathcal{X}|}$ and costs $\mathrm{C} \in R_{\geq 0}^{n \times m \times |\mathcal{X}|}$.

GLOBE-CE algorithm

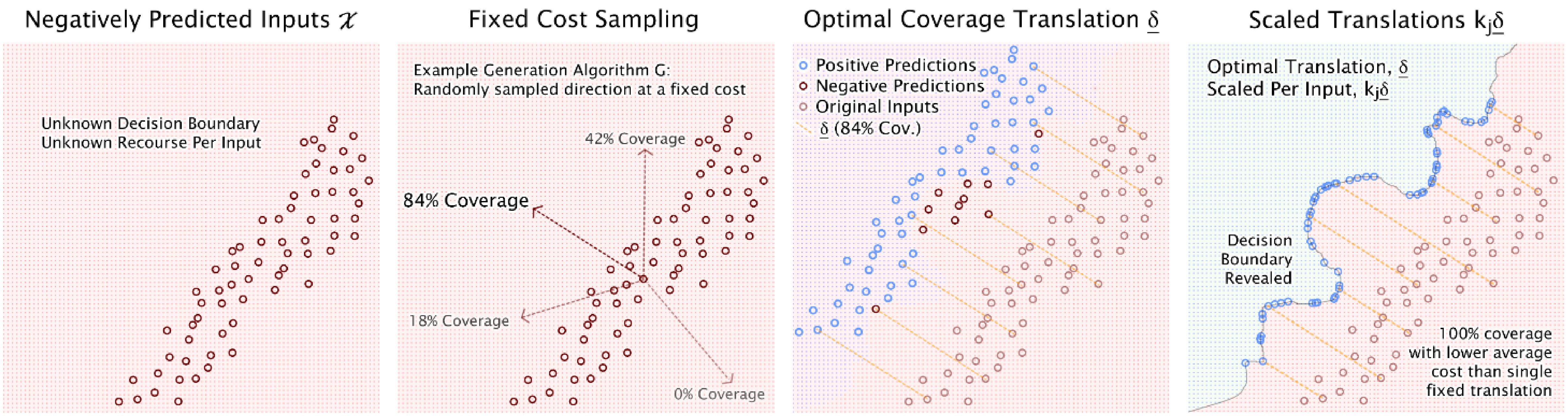


Figure 3. The GLOBE-CE framework (Algorithm 1) for an example generation algorithm G. Cost is ℓ_2 distance. **Left:** Negative predictions, $\mathcal{X}$. **Left Center:** We sample translations at a fixed cost, computing the coverage of each translation. **Right Center:** The translation with highest coverage is selected. **Right:** We scale $\underline{\delta}$ per input, returning the k_j value required for each input, where j indexes a vector of scalars $\underline{k}$. Theorems 4.1 and 4.2 bridge the gap between scaling translations and the discontinuous nature of categorical features.

GLOBE-CE algorithm

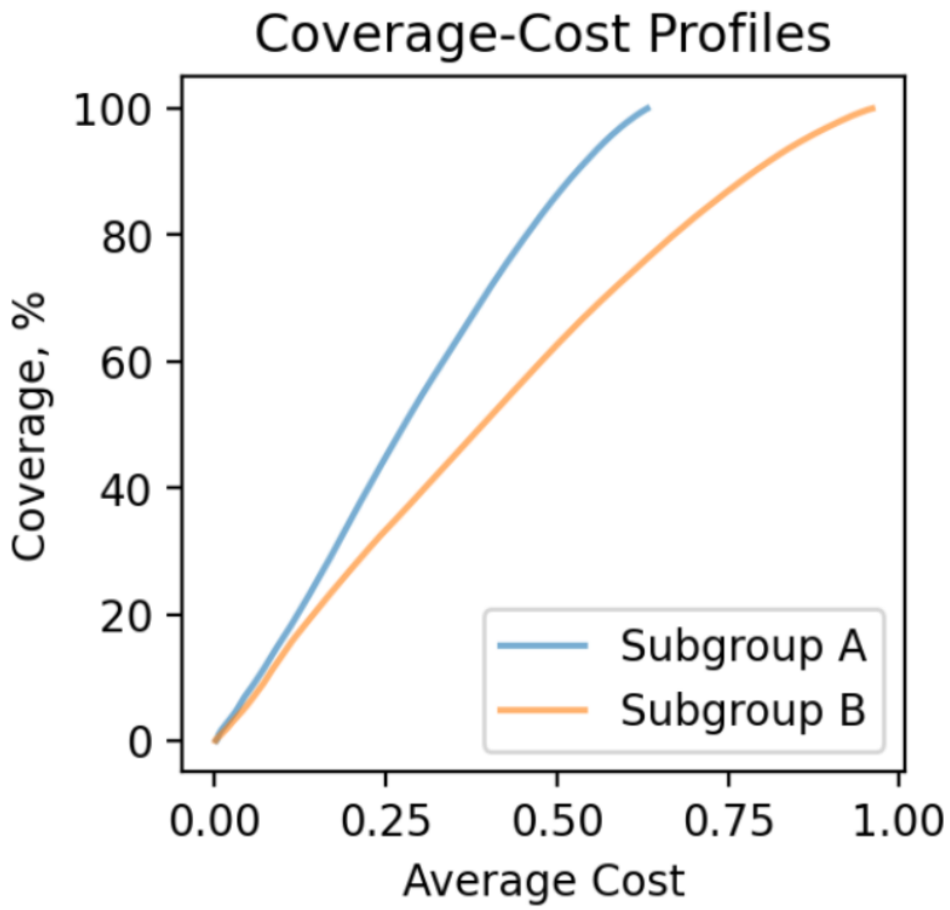
Algorithm 1 GLOBE-CE Framework

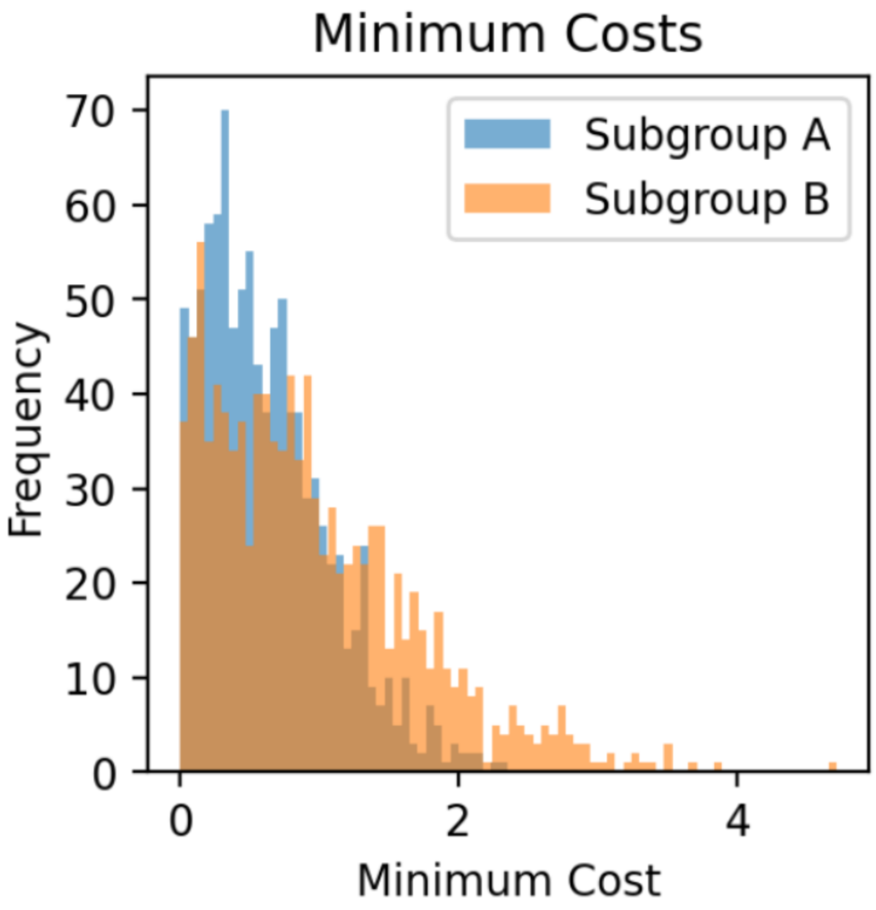
Input: B, $\mathcal{X}$, G, n, $\underline{k}$, $cost$

1: $\Delta = G(B, \mathcal{X}, n)$ ▷ Generate GCE Directions
2: **for** $1 \leq i \leq n$ **do** ▷ For all GCEs
3:     **for** $1 \leq j \leq |\underline{k}|$ **do** ▷ For all Scalars
4:         $\mathcal{X}'_{ij} = round(\mathcal{X} + k_j \underline{\delta}_i)$ ▷ Counterfactuals
5:         $\mathcal{Y}'_{ij} = B(\mathcal{X}'_{ij})$ ▷ Predictions
6:         $\mathcal{C}_{ij} = cost(\mathcal{X}, \mathcal{X}'_{ij})$ ▷ Costs
7:     **end for**
8: **end for**

Output: Counterfactuals $\mathcal{X}'$, Predictions $\mathcal{Y}'$, Costs $\mathcal{C}$ (For all Inputs $\mathcal{X}$, Translations Δ and Scalars $\underline{k}$)

GLOBE-CE: Interpreting translation directions





GLOBE-CE: Cumulative Rules Chart

Table 2. Example *Cumulative Rules Chart (CRC)* for categorical features in the German Credit dataset, representing the optimal GLOBE-CE translation at 5 scalar values. Rules are cumulatively added (from top to bottom), resulting in an increase in coverage and cost.

Feature(s)	**New Rule Added**	**New Inputs**		**All Inputs**	
		Coverage	**Cost**	**Coverage**	**Cost**
Account Status	If F2, Then F4	+33.5%	1.00	33.5%	1.00
Account Status	If F3, Then F4	+2.5%	1.00	36.0%	1.00
Account Status	If F1, Then F4	+45.2%	1.00	81.2%	1.00
Telephone	If F2, Then F1	+2.5%	1.80	83.7%	1.02
Employment	If Not F4, Then F4	+10.2%	1.95	93.9%	1.12

Experiments setup

- Models:
  - Deep Neural Network
  - XGBoost
  - Logistic Regression
- Datasets:
  - COMPAS (recidivism)
  - German Credit (credit risk)
  - Default Credit (payment defaults)
  - HELOC (credit risk)
- Specific generation algorithm $G(B, \mathcal{X}, n, n_s, c, n_f, p)$ – uniform sampling of n_s translations at a fixed cost c with randomly chosen features n_f and the power p to which random samples between 0 and 1 are raised

Baseline - AReS

Table 1. Comparison of the AReS and GLOBE-CE algorithms, highlighting differences in methodology, feature handling, performance, and efficiency. The main differences include the handling of continuous features as well as the overall efficiency of both methods.

Comparison	**AReS**	**GLOBE-CE**
Algorithm	Generates hundreds/thousands of items $\mathcal{SD}$ Searches $\mathcal{SD}^3$ for valid triples, V Optimises V to select a smaller set of triples, R	Generates n GCE directions Scales each direction across all inputs Returns information on minimum cost per input
Continuous Features	Bins continuous features, displayed as If-Then rules (searches for combinations between commonly occurring bins)	Does not bin continuous features, displayed as addition/subtraction (no binning leads to performance improvements)
Categorical Features	Displayed as If-Then rules	We prove that (scaled) translations can also be expressed as If-Then rules
Performance	Lower coverage and higher average cost	Higher coverage and lower average cost
Efficiency	Computationally slow (hours for best performance)	Computationally fast (seconds)

Experiments results

Table 3. Evaluating the reliability (coverage/cost) and efficiency of GLOBE-CE against AReS. Highlighted in red are GCEs that a) achieve below 10% coverage or b) require computation time in excess of 10,000 seconds (≈3 hours). Best metrics are shown in **bold**.

Models	Algorithms	COMPAS			German Credit			Default Credit			HELOC		
		Cov.	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*
DNN	AReS	51%	2.31	101s	73%	1.6	2712s	7.22%	1.0	7984s	5.4%	1.0	9999s
	Fast AReS	64%	**1.45**	32.0s	72%	1.43	12.8s	99.8%	4.2	37.3s	52%	5.5	109.1s
	GLOBE-CE	66%	1.53	**7.08s**	85%	1.2	**2.28s**	98.5%	1.3	**3.6s**	93%	4.3	**4.66s**
	dGLOBE-CE	**70%**	1.46	9.15s	**90%**	**1.1**	2.63s	**100%**	**1.1**	7.86s	**95%**	3.8	5.46s
XGB	AReS	45%	1.9	205s	61%	1.5	2092s	11%	1.0	9999s	1.7%	1.0	9999s
	Fast AReS	83%	1.9	47.6s	65%	1.75	34.33s	93%	2.3	29.97s	28%	**2.1**	93.58s
	GLOBE-CE	78%	1.8	**9.61s**	**95%**	**1.02**	**5.04s**	96%	1.1	**2.94s**	58%	2.4	**4.7s**
	dGLOBE-CE	**91%**	**1.4**	12.4s	83%	1.03	5.95s	**100%**	**0.7**	6.35s	**80%**	2.4	5.6s
LR	AReS	79%	1.5	506s	85%	1.3	3566s	31%	1.2	9999s	4.8%	1.0	9999s
	Fast AReS	82%	1.7	43.0s	85%	1.3	9.3s	99%	2.1	17.82s	92%	1.6	127.3s
	GLOBE-CE	83%	1.20	**8.43s**	82%	**1.2**	**3.39s**	**100%**	**1.0**	**3.42s**	**100%**	**0.5**	**3.11s**
	dGLOBE-CE	**84%**	**1.18**	11.7s	**91%**	1.3	3.87s	**100%**	**1.0**	7.21s	**100%**	**0.5**	3.85s

User study

- User study was performed to analyse and compare the efficacy of GLOBE-CE and AReS in detecting recourse biases
- 24 participants, all with a background in AI and ML
- The study utilises two “black box” models:
  - decision tree with a model bias against females, though with a recourse bias exhibited against males due to the nature of the data distribution
  - SVM with a recourse bias against a *ForeignWorker* subgroup

User study

If Sex = Male:

If Job = No **and** Property = No,
Then Job = Yes **and** Property = Yes

If Healthcare = No,
Then Healthcare = Yes

If Sex = Female:

If Job = No **and** Property = No **and** Savings = No,
Then Job = Yes **and** Property = Yes **and** Savings = Yes

If Healthcare = No,
Then Healthcare = Yes

Figure 5. Depiction of Black Box 1, with *model* bias against females, yet *recourse* bias against males. 90% of rejected females satisfy the first rule with cost 3, and require healthcare with cost 1. In contrast, 90% of rejected males have healthcare, but require the first rule with cost 2, resulting in higher average recourse costs.

User study

- For each explanation, the user study asks two questions
  - do you think there exists bias in the presented recourse rules?
  - explain the reasoning behind your choice.

Table 4. Bias detection results from user studies. Bias and correct columns: number of users that identified a bias and number of users that described it correctly, respectively.

User Studies Breakdown	AReS *Bias?*	AReS *Correct?*	GLOBE-CE *Bias?*	GLOBE-CE *Correct?*
Black Box 1	7/8	0/8	7/8	7/8
Black Box 2	1/8	0/8	5/8	4/8

Conclusion

- This work proposes GLOBE-CE, a novel GCE framework that further improves on the issues faced by the current SOTA and addresses the issues associated with prior work:
  - requiring GCEs to be fixed-magnitude translations
  - computational complexity
- Experiments with four public datasets and user studies demonstrate the efficacy of our proposed framework in generating accurate global explanations that assist in identifying recourse biases

Questions?