Learning Certified Individually Fair Representations

Anian Ruoss, Mislav Balunović, Marc Fischer, Martin Vechev

Fair Representation Learning

We consider individual fairness, which requires that similar individuals should be treated similarly. More concretely, we focus on individual fairness in the following context [McNamara et al., 2019]:

Data Regulator: defines fairness notion for the task
Data Producer: transforms user data into latent representation
Data Consumer: performs predictions based on latent representation

Fair Representation Learning

We consider individual fairness, which requires that similar individuals should be treated similarly. More concretely, we focus on individual fairness in the following context [McNamara et al., 2019]:

Data Regulator: defines fairness notion for the task
Data Producer: transforms user data into latent representation
Data Consumer: performs predictions based on latent representation

Key Challenge: enforce individual fairness in the setting described above by learning and certifying individually fair representations

Data Regulator

We propose an individual fairness notion in terms of interpretable logical constraints. For example, for an individual $x \in R^N$ we want all similar individuals, defined as $x' \in R^N$ such that

$$\phi(x, x') := \bigwedge_{i \in Categorical \setminus \{race,\, gender\}} (x_i = x_i') \bigwedge_{j \in Numerical} |x_j - x_j'| \leq \alpha,$$

to be classified the same.

Individual Fairness

We define $S_\phi(x) := \{x' \in R^N \mid \phi(x, x')\}$ to be the set of all individuals similar to x. We want to **certify** for individual x in the test dataset,

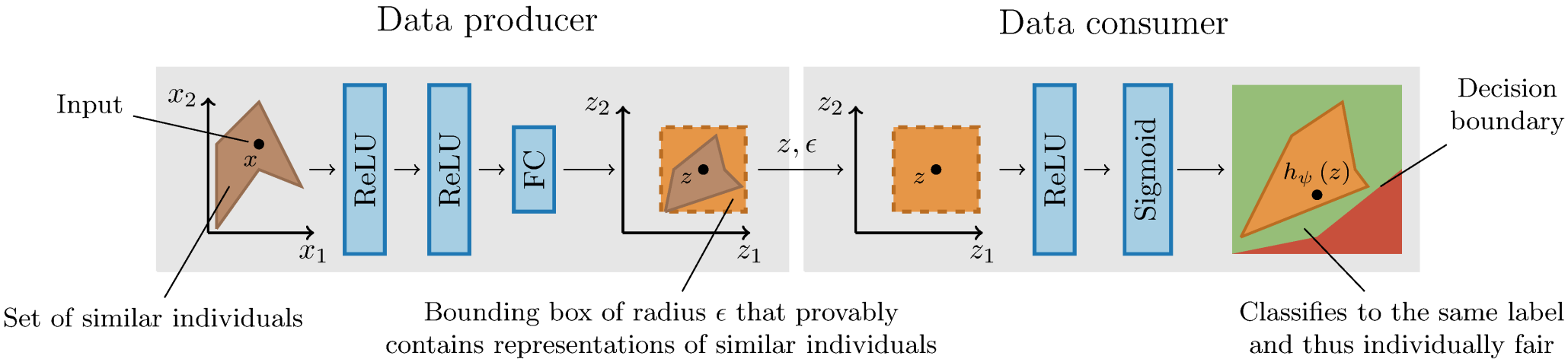
$$\forall x' \in S_\phi(x) \Longrightarrow \mu\big(M(x), M(x')\big),$$

where $M : R^N \to R^O$ is the model, and $\phi : R^N \times R^N \to \{0, 1\}$ and $\mu : R^O \times R^O \to \{0, 1\}$ are binary similarity measures on inputs and outputs.

We consider classification with $\mu\big(M(x), M(x')\big) \Leftrightarrow M(x) = M(x')$.

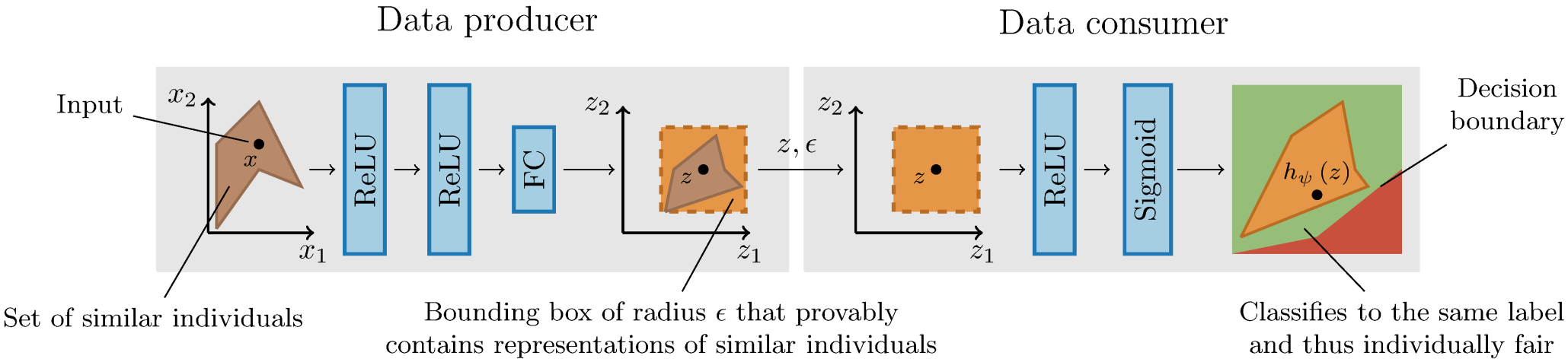
Data Producer

- trains encoder $f_\theta : R^N \rightarrow R^K$, where K is the dimension of the latent space, such that $\forall x' \in S_\phi(x) \Longrightarrow |f_\theta(x) - f_\theta(x')|_\infty \leq \delta$
- encodes $S_\phi(x)$ and f_θ as mixed-integer linear program (MILP) to compute ϵ such that $f_\theta\big(S_\theta(X)\big) \subseteq \{z' || z - z'|_\infty \leq \epsilon\}$ for $z := f_\theta(x)$



Data Consumer

- obtains z and ϵ from data producer and uses standard local robustness training to obtain a classifier $h_\psi : R^K \rightarrow R^O$ that is robust to l_∞-perturbations of magnitude ϵ around z
- uses neural network robustness verifier to certify ϵ-robustness for z



Experimental Evaluation

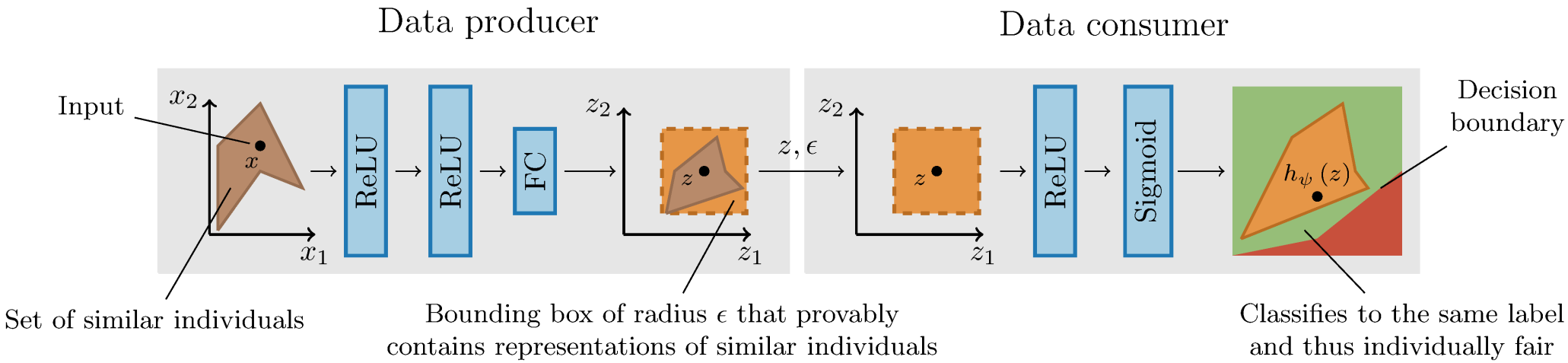
$$\phi(x, x') := \bigwedge_{i \in Categorical \setminus \{race, gender\}} (x_i = x_i') \bigwedge_{j \in Numerical} |x_j - x_j'| \leq \alpha,$$

	Accuracy (%)		Certified Individual Fairness (%)	
Dataset	**Baseline**	**LCIFR**	**Baseline**	**LCIFR**
Adult	83.3	81.3	47.5	97.6
Compas	65.6	63.7	30.9	75.6
Crime	84.4	81.5	6.2	63.3
German	76.0	70.0	68.0	95.5
Health	80.7	80.7	24.7	97.3
Law School	84.4	84.5	11.6	28.9

Learning certified individually fair representations (LCIFR) significantly increases the percentage of points with certified individual fairness, without compromising accuracy.

Conclusion

$$\phi(x, x') := \bigwedge_{i \in Categorical \setminus \{race, gender\}} (x_i = x_i') \bigwedge_{j \in Numerical} |x_j - x_j'| \leq \alpha,$$



Code: https://github.com/eth-sri/lcifr