ECCV TEL AVIV 2022

Latent Space Smoothing for Individually Fair Representations

Momchil Peychev

Anian Ruoss

Mislav Balunović

Maximilian Baader

Martin Vechev

ETH zürich

SRILAB

DeepMind

Individually Fair Representation Learning

Individual fairness: requires similar individuals to be treated similarly.

Individual fairness in the context of:

Data Regulator: defines a fairness notion for the task.
Data Producer: learns a fair representation that encodes the data.
Data Consumers: employ the transformed data to make predictions.

Prior work: *enforces* and *certifies* individual fairness for low-dimensional tabular data.

This work: scaling to high-dimensional data and real-world models.

McNamara et al., Costs and Benefits of Fair Representation Learning, AIES 2019
Ruoss et al., Learning Certified Individually Fair Representations, NeurIPS 2020

Data Regulator

For a given person, all people differing only in skin tone should be classified the same.

Key challenge: high-level semantic attributes cannot be captured conveniently in the input space of $\boldsymbol{x} \in \mathbb{R}^n$.

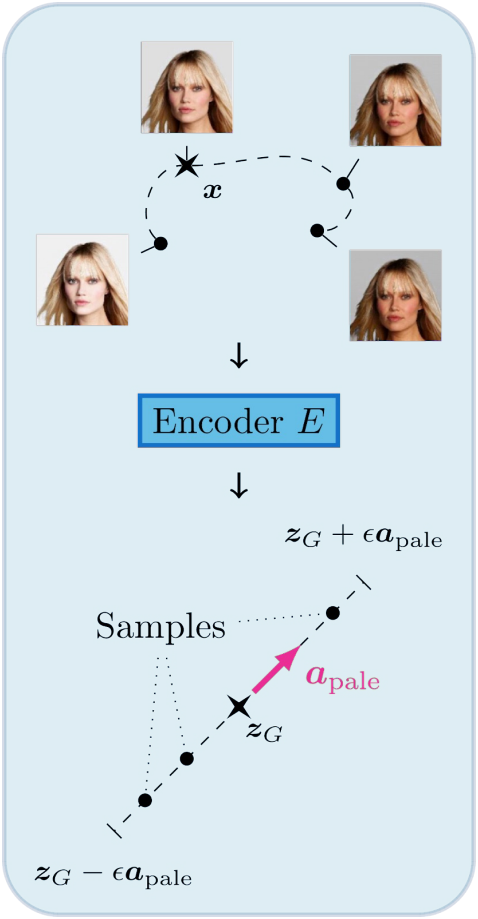
Leverage Glow $G = (E, D)$ and compute attribute vector $\boldsymbol{a} \in \mathbb{R}^q$ in the latent space of G.

Similarity set for $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{z}_G = E(\boldsymbol{x})$:

$$S(\boldsymbol{x}) \coloneqq \{\boldsymbol{z}_G + t \cdot \boldsymbol{a} \mid |t| \leq \epsilon\} \subseteq \mathbb{R}^q.$$

Goal: for an individual $\boldsymbol{x}$ in the test set, **certify**

$$\forall \boldsymbol{z} \in S(\boldsymbol{x}): \hat{C} \circ \hat{R}(\boldsymbol{z}_G) = \hat{C} \circ \hat{R}(\boldsymbol{z}).$$



Data Producer

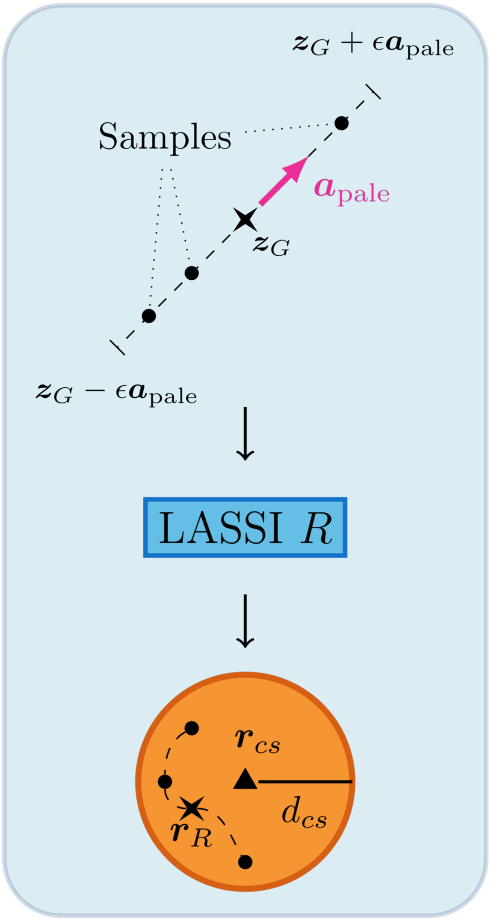
$$S(\boldsymbol{x}) \coloneqq \{\boldsymbol{z}_G + t \cdot \boldsymbol{a} \mid |t| \leq \epsilon\}$$

- Learn a representation $R: \mathbb{R}^q \rightarrow \mathbb{R}^k$ that maps similar individuals close together.

  Adversarial loss: $\mathcal{L}_{adv}(\boldsymbol{x}) = \max_{\boldsymbol{z} \in S(\boldsymbol{x})} \|R(\boldsymbol{z}_G) - R(\boldsymbol{z})\|_2$.

- Adversarial training → no formal guarantees.

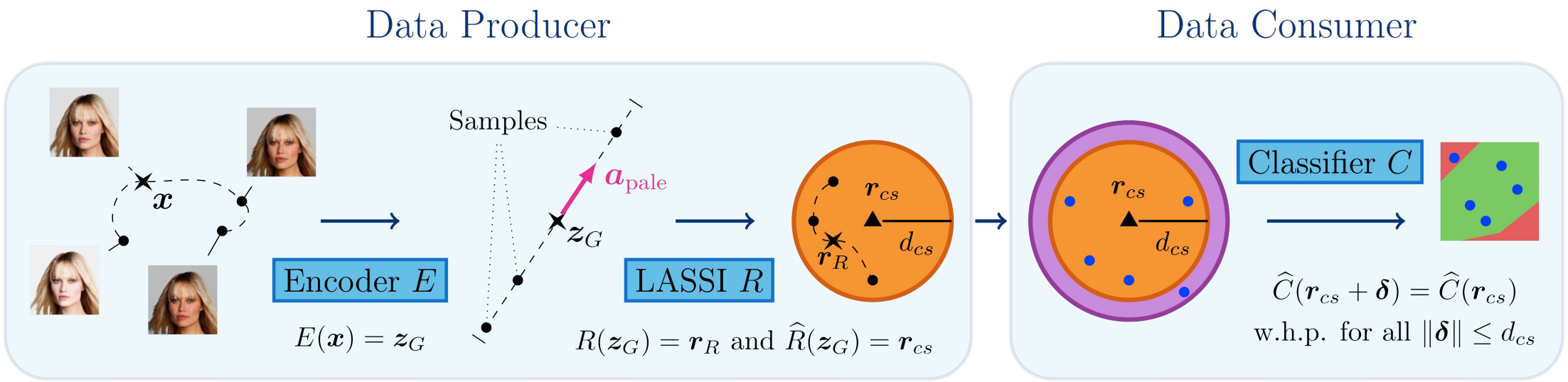
- Center smoothing on R, $\hat{R}(\boldsymbol{z}_G)$, produces a center $\boldsymbol{r}_{cs}$ and radius d_{cs} such that $\forall \boldsymbol{z} \in S(\boldsymbol{x})$: $\left\|\boldsymbol{r}_{cs} - \hat{R}(\boldsymbol{z})\right\|_2 \leq d_{cs}$ with high probability.



Kumar and Goldstein, Center Smoothing: Certified Robustness for Networks with Structured Outputs, NeurIPS 2021

Data Consumer

- Randomized smoothing on the downstream classifier $C: \mathbb{R}^k \to \mathcal{Y}$, $\hat{C}(\boldsymbol{r}_{cs})$, to obtain its ℓ_2-robustness radius d_{rs} around $\boldsymbol{r}_{cs}$.
- End-to-end certificate: if $d_{cs} < d_{rs}$, then the end-to-end model provably satisfies individual fairness at $\boldsymbol{x}$ with high probability.



Cohen et al., Certified Adversarial Robustness via Randomized Smoothing, ICML 2019

Evaluation

Pale Skin

Young

Blond Hair

Heavy Makeup

Pale Skin + Young

Pale Skin + Young + Blond Hair

Glow reconstructions corresponding to points from the similarity sets for multiple sensitive attribute combinations from the CelebA dataset.

Evaluation

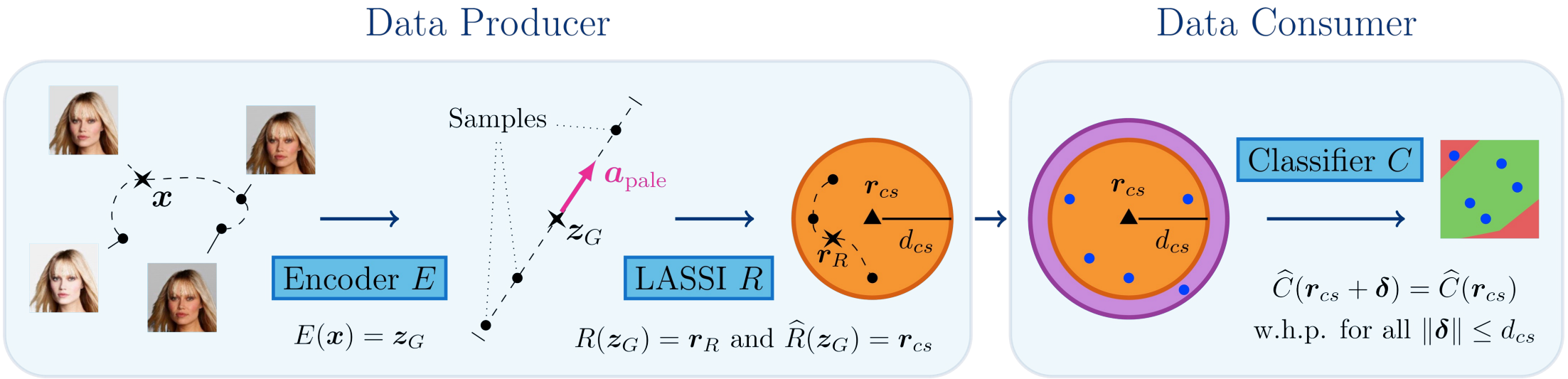
		Naive		Data Augmentation		LASSI (ours)	
Task	Sensitive attribute(s)	Acc	Fair	Acc	Fair	Acc	Fair
Smiling	Pale Skin	**86.3**	0.6	85.7	12.2	85.9	**98.0**
	Young	**86.3**	38.2	85.9	43.0	**86.3**	**98.8**
	Blond Hair	86.3	3.4	**86.6**	9.4	86.4	**94.7**
	Heavy Makeup	**86.3**	0.4	85.3	13.7	85.6	**91.3**
	Pale Skin + Young	**86.0**	0.4	85.8	9.9	85.8	**97.3**
	Pale + Young + Blond	86.2	0.0	**86.4**	3.6	85.5	**86.5**

Our method (LASSI) significantly increases the percentage of points for which we can certify individual fairness, without affecting the classification accuracy.

Check our paper for further details

- Attribute vector computation
- Transfer learning experiments
- Experiments on the FairFace dataset
- More examples of Glow reconstructions of similar individuals

Conclusion



Code and pretrained models: https://github.com/eth-sri/lassi