Latent Space Smoothing for Individually Fair Representations

Momchil Peychev, Anian Ruoss, Mislav Balunović, Maximilian Baader, Martin Vechev

ETH zürich · DeepMind · ECCV TEL AVIV 2022

https://github.com/eth-sri/lassi

Individual Fairness for High-Dimensional Data

Problem Setting

Individual fairness: treat similar individuals similarly. In the context of:

- **Data Regulator:** defines a fairness notion for the task
- **Data Producer:** learns a fair representation of the data
- **Data Consumers:** make predictions from the representation

This work:

i. Design a suitable input similarity metric
ii. Enforce that similar individuals are *provably* treated similarly

for *high-dimensional data* and *real-world models*.

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 the original data $\boldsymbol{x} \in \mathbb{R}^n$.

Key Insight: Similarity via a Generative Model

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

Similarity set

- In the latent space of G: $S(\boldsymbol{x}) \coloneqq \{\boldsymbol{z}_G + t \cdot \boldsymbol{a} \mid |t| \leq \epsilon\} \subseteq \mathbb{R}^q$
- In the input space $\mathbb{R}^n$: $S^{\text{in}}(\boldsymbol{x}) \coloneqq D(S(\boldsymbol{x})) \subseteq \mathbb{R}^n$

Goal: construct an end-to-end model $M: \mathbb{R}^n \to \mathcal{Y}$ for which, given an input $\boldsymbol{x}$, we can **certify** individual fairness of the form

$$\forall \boldsymbol{x}' \in S^{\text{in}}(\boldsymbol{x}): M(\boldsymbol{x}) = M(\boldsymbol{x}')$$

Latent Space Smoothing for Individually Fair Representations (LASSI)

Data Producer

1. Glow encoder E and a representation $R: \mathbb{R}^q \to \mathbb{R}^k$ trained with:
Adversarial loss: map similar individuals close together in $\mathbb{R}^k$

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

Classification loss: ensure utility for downstream tasks

$$\mathcal{L}_{cls}(\boldsymbol{x}, y) = \text{cross\_entropy}(C_{aux} \circ R(\boldsymbol{z}_G), y)$$

Reconstruction loss: for transfer learning, preserve original signal

$$\mathcal{L}_{recon}(\boldsymbol{x}) = \left\|\boldsymbol{z}_G - Q\big(R(\boldsymbol{z}_G)\big)\right\|_2 \text{ with } Q: \mathbb{R}^k \to \mathbb{R}^q$$

R, C_{aux} and Q are trained jointly, trading off fairness, accuracy and transferability: $L = \lambda_1 \mathcal{L}_{cls}(\boldsymbol{x}, y) + \lambda_2 \mathcal{L}_{adv}(\boldsymbol{x}) + \lambda_3 \mathcal{L}_{recon}(\boldsymbol{x})$.

2. Adversarial training → no formal guarantees. Center smoothing on R, $\hat{R}(\boldsymbol{z}_G)$, provably mapping similar individuals close together:

$$\boldsymbol{r}_{cs}, d_{cs} \leftarrow \hat{R}(\boldsymbol{z}_G) \text{ and } \forall \boldsymbol{z} \in S(\boldsymbol{x}): \left\|\boldsymbol{r}_{cs} - \hat{R}(\boldsymbol{z})\right\|_2 \leq d_{cs}$$

with probability at least $1 - \alpha_{cs}$ [1].

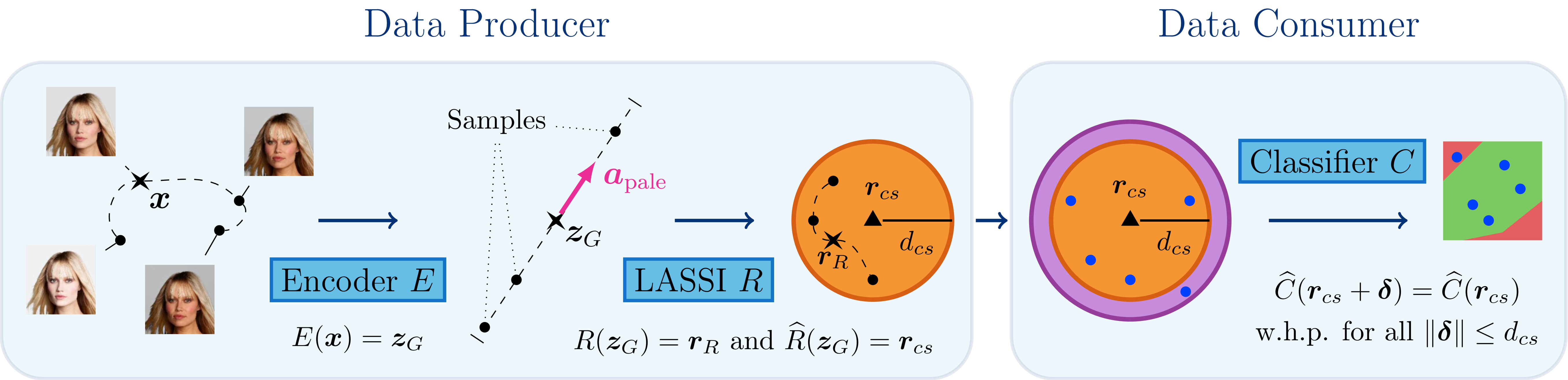
Data Consumer

Randomized smoothing [2] (with confidence α_{rs}) 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 $\boldsymbol{r}_{cs}$. The classifier C is trained after training R is completed.

End-to-end Fairness Certificate for $M = \hat{C} \circ \hat{R} \circ E$

Given input $\boldsymbol{x} \in \mathbb{R}^n$, let $\boldsymbol{z}_G = E(\boldsymbol{x})$

1. $\boldsymbol{r}_{cs}, d_{cs} \leftarrow \hat{R}(\boldsymbol{z}_G)$ and $d_{rs} \leftarrow \hat{C}(\boldsymbol{r}_{cs})$
2. If $d_{cs} < d_{rs}$, then ***provably*** $\forall \boldsymbol{x}' \in S^{\text{in}}(\boldsymbol{x}): M(\boldsymbol{x}) = M(\boldsymbol{x}')$ with probability at least $1 - \alpha_{cs} - \alpha_{rs}$ for $M = \hat{C} \circ \hat{R} \circ E$.



Experimental Evaluation of LASSI

Datasets: CelebA 64x64 (and FairFace – see paper), 312 test samples
Similarity set: $\boldsymbol{a} = \boldsymbol{z}_{G,pos} - \boldsymbol{z}_{G,neg}$ [3], $\epsilon = 1$
Naive: $\lambda_1 = 1, \lambda_2 = \lambda_2 = 0$, LASSI: $\lambda_2 = 0.05$, Transfer: $\lambda_1 = 0, \lambda_3 = 0.1$

Single & Multiple Sensitive Attributes

Task	Sensitive attribute(s)	Naive Acc	Naive Fair	DataAug Acc	DataAug Fair	LASSI Acc	LASSI 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**

LASSI significantly increases the percentage of points for which we can certify individual fairness, without affecting the accuracy.

Transfer Learning

Sens. attrib.: / Transfer task	Pale Skin Acc	Pale Skin Fair	Pale + Young Acc	Pale + Young Fair	Pale+Young+Blond Acc	Pale+Young+Blond Fair
Smiling	86.2	93.1	85.9	92.2	85.1	87.0
HighCheeks	81.7	92.6	80.8	93.0	80.6	84.5
MouthOpen	81.5	91.2	81.6	90.1	82.5	80.8
Lipstick	88.3	94.0	85.1	90.6	86.2	81.0
HeavyMakeup	86.5	93.0	83.7	90.0	83.3	80.4
Wavy Hair	79.2	93.3	77.6	91.5	78.8	85.3
Eyebrows	78.3	92.1	77.8	92.2	78.7	85.6

The method is modular: LASSI can still achieve high certified individual fairness even when downstream tasks are not known.

Computing $\boldsymbol{a}$

LASSI is ***independent*** of the actual computation of $\boldsymbol{a}$. We instantiate it with four different attribute vector types in our paper [3,4,5,6].

References:

1. Kumar and Goldstein, Center Smoothing: Certified Robustness for Networks with Structured Outputs — NeurIPS 2021
2. Cohen et al., Certified Adversarial Robustness via Randomized Smoothing — ICML 2019
3. Kingma and Dhariwal, Glow: Generative flow with invertible 1x1 convolutions — NeurIPS 2018
4. Denton et al., Detecting bias with generative counterfactual face attribute augmentation — arXiv 2019
5. Ramaswamy et al., Fair attribute classification through latent space debiasing — CVPR 2021
6. Li and Xu, Discover the unknown biased attribute of an image classifier — ICCV 2021