Latent Space Smoothing for Individually Fair Representations (LASSI)

Data Producer

1. Glow encoder $E$ and a representation $R: \mathbb{R}^d \rightarrow \mathbb{R}^k$ trained with:
   
   **Adversarial loss**: map similar individuals close together in $\mathbb{R}^k$
   
   $$L_{adv}(x) = \max_{z \in S(x)} \| R(z) - R(x) \|_2$$

Classification loss: ensure utility for downstream tasks

$$L_{cls}(x, y) = \text{cross entropy}(C_{aux} \circ R(z), y)$$

Reconstruction loss: for transfer learning, preserve original signal

$$L_{recon}(x) = \| z_g - Q(R(z_g)) \|_2$$

$C_{aux}$ and $Q$ are trained jointly, trading off fairness, accuracy and transferability:

$$L = \lambda_1 L_{cls}(x, y) + \lambda_2 L_{adv}(x) + \lambda_3 L_{recon}(x).$$

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

$$r_{cs}, d_{cs} \leftarrow \hat{R}(x_g)$$

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

Data Consumer

Randomized smoothing [2] (with confidence $\alpha_{cs}$) on the downstream classifier $C: \mathbb{R}^k \rightarrow \mathbb{y}, \hat{C}(r_{cs})$, to obtain its $\ell_2$-robustness radius $d_{cs}$ around $r_{cs}$. The classifier $\hat{C}$ is trained after training $R$ is completed.

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

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

1. $r_{cs}, d_{cs} \leftarrow \hat{R}(z_g)$ and $d_{cs} \leftarrow \hat{C}(r_{cs})$

2. If $d_{cs} < \delta_{rs}$ then **provably** $\forall x' \in S_{in}(x); M(x) = M(x')$ with probability at least $1 - \alpha_{cs} - \alpha_{rs}$ for $M = \hat{C} \circ \hat{R} \circ E$.

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 ICLR 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