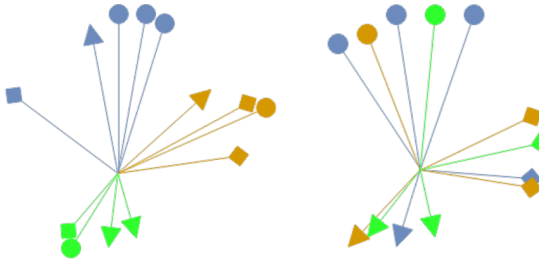
Debiasing through alignment and disentanglement: EnD

Align all bias-conflicting samples and repel bias-aligned positives

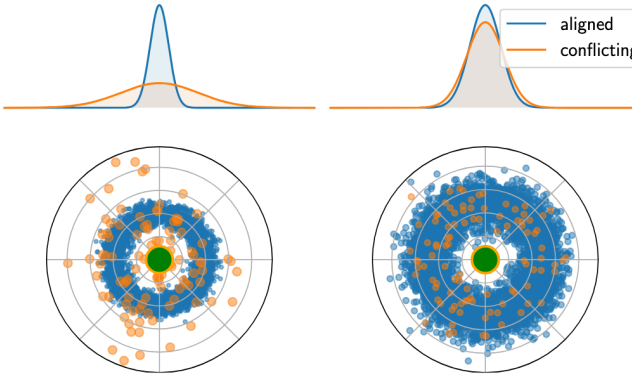
$$\mathcal{R}_i^{end} = \alpha \sum_{a \in B(i)} |z_i \cdot z_a| - \beta \sum_{j \in J(i)} z_i \cdot z_j .$$



(a) → (b)

Moment matching between bias-conflicting and bias-aligned positives: **FairKL**

$$\mathcal{R}^{FairKL} = D_{KL}(B^{align} || B^{confl})$$



Debiasing without bias labels: **Unsupervised Debiasing**