Bias-inducing geometries: exactly solvable data model with fairness implications

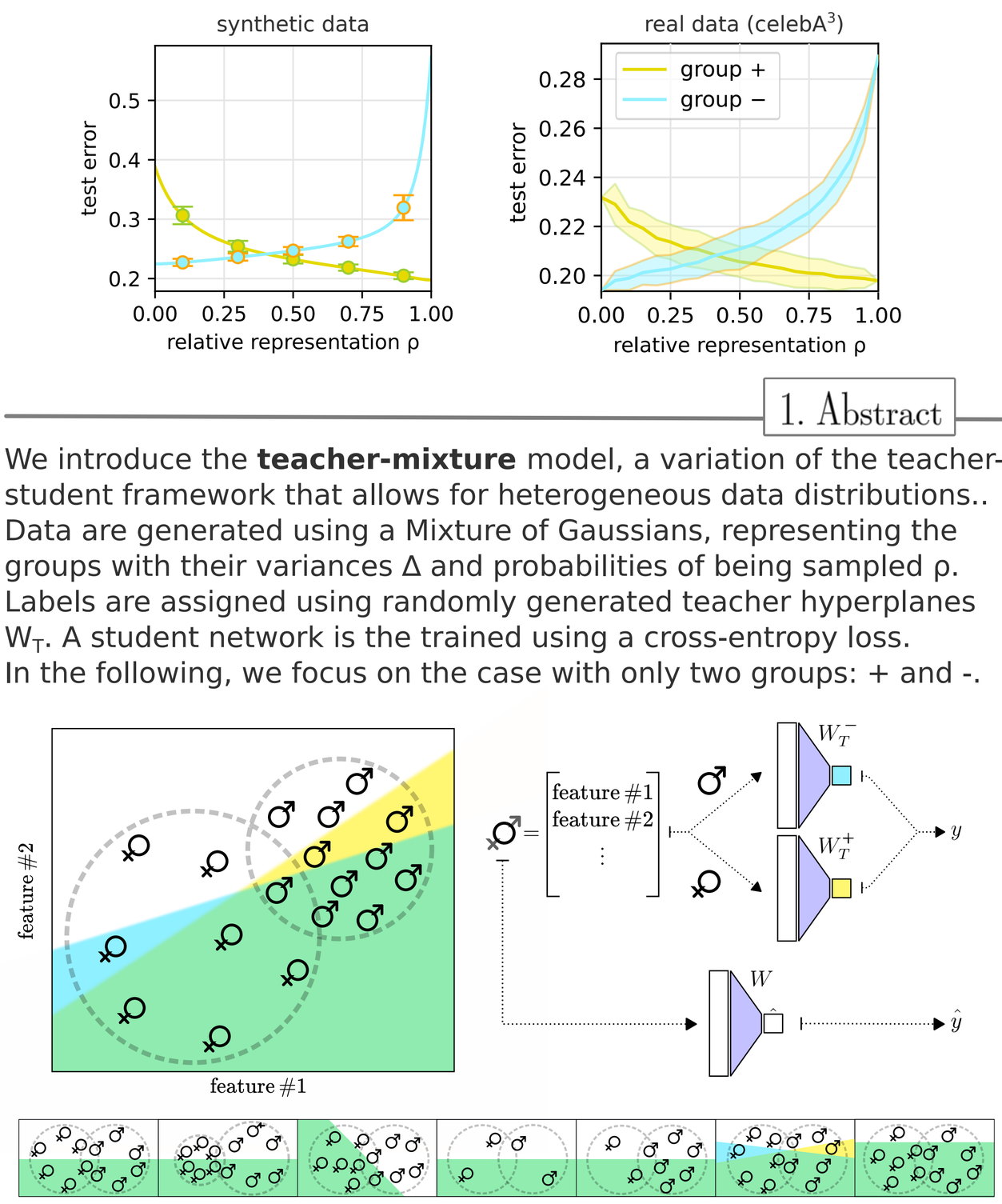
Stefano **Sarao Mannelli**[1], Federica **Gerace**[2], Negar **Rostamzadeh**[3], Luca **Saglietti**[4]

[1]Gatsby & SWC, UCL. [2]SISSA. [3]Google Research. [4]Computing Sciences department, Bocconi.

TL;DR classifiaction bias can emerge as a consequence of geometrical properties even in supposedely fair settings.

1. Abstract

We introduce the **teacher-mixture** model, a variation of the teacher-student framework that allows for heterogeneous data distributions.. Data are generated using a Mixture of Gaussians, representing the groups with their variances Δ and probabilities of being sampled ρ. Labels are assigned using randomly generated teacher hyperplanes W_T. A student network is the trained using a cross-entropy loss.
In the following, we focus on the case with only two groups: + and -.



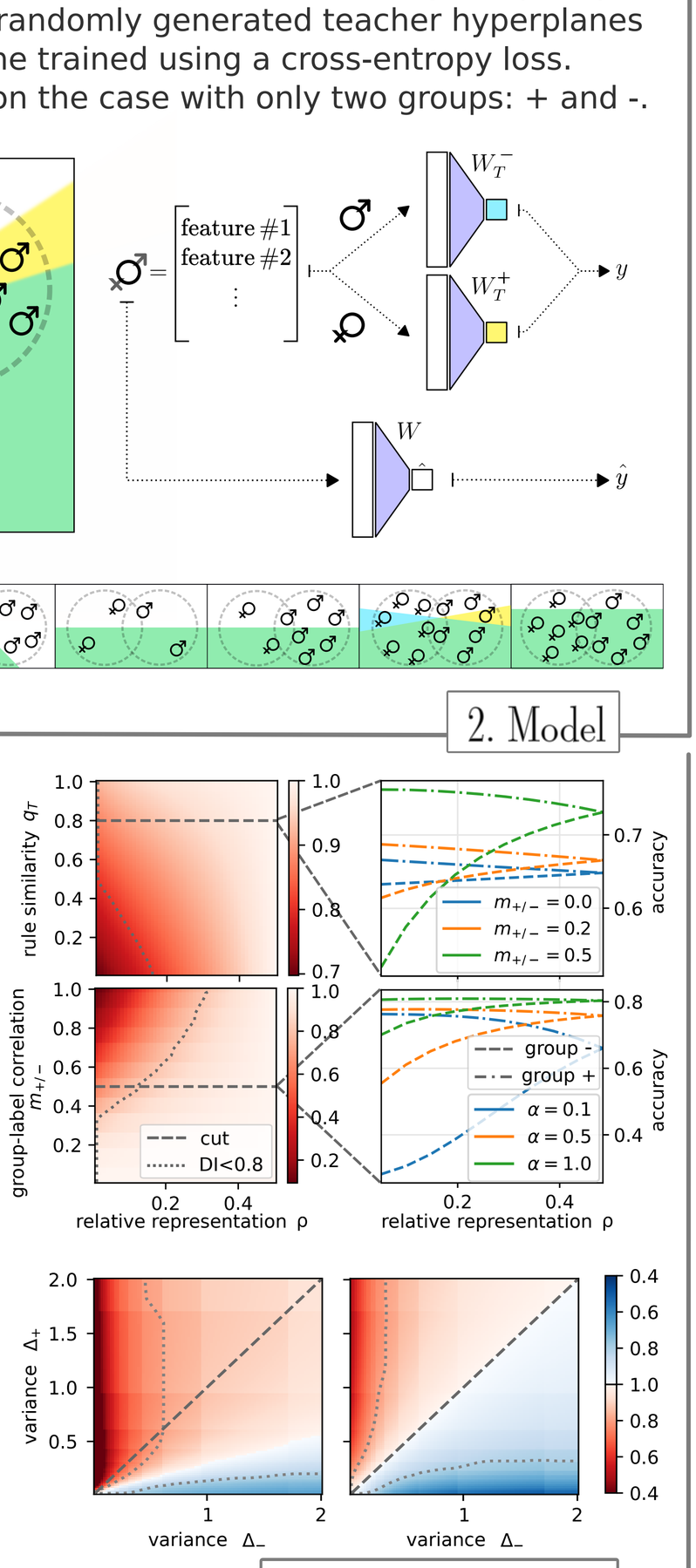
2. Model

Using disparate impact

$$DI = \frac{p(\hat{y} = y|+)}{p(\hat{y} = y|-)}$$

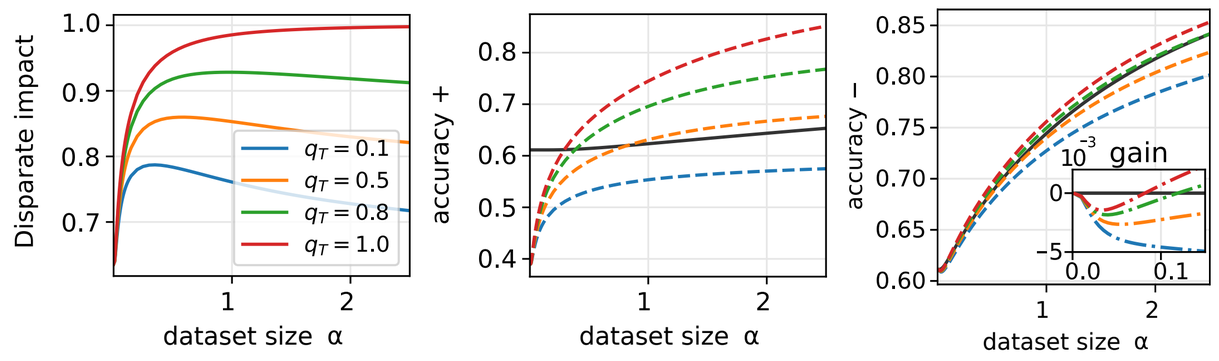
as a measure for bias, the heatmaps show the role of the different data properties on bias, with darker colours indicating stronger bias.

We found strong bias in model mistmatched case (q_T<1) when a student tries to fit two teachers, but also in the model matched case (qT=1) due to different sampling probabilities or different level of variance in the two groups.



3. Emergence of bias

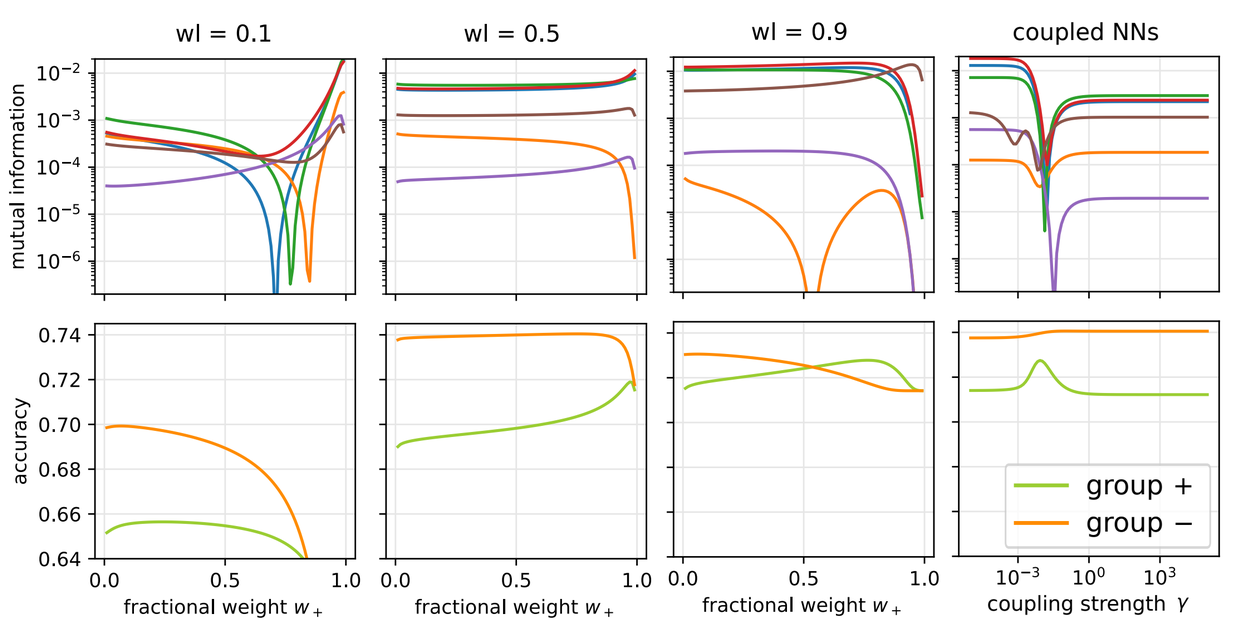
Assuming that the groups are known, it is tempting to split the dataset and train different classifiers for each group.
However, this can lead to a worse performance. Joint training is beneficial in the regimes where the dataset are neither very small nor very large. This Goldilock regime increases with more aligned teacher hyperplanes ($q_T \to 1$). Joint training allows to exploit the task similarity, akin to the positive transfer effect in transfer learning[4].



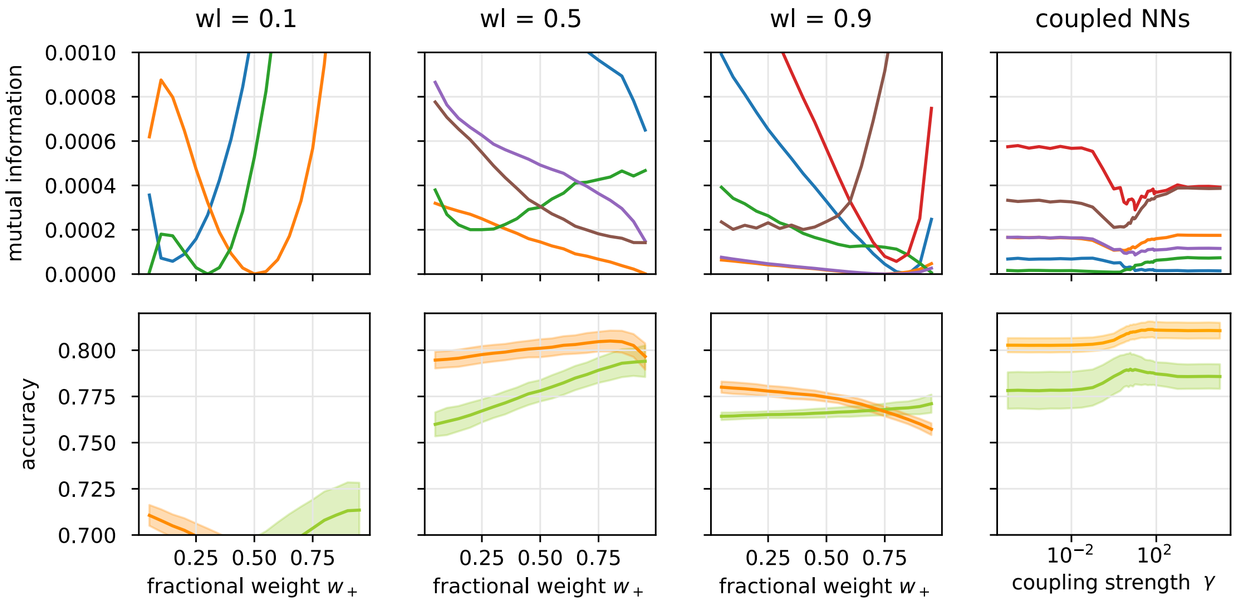
4. Positive transfer

So far we considered only the notion of disparate impact to measure bias, however several metrics including statistical parity, equal opportunities, equal accuracy, equal odds, predicted parity 1, and predicted parity 10, are used in the literature. These metrics want an event -e.g. estimated labels=correct labels- to be independent with respect to the group membership. Using mutual information we can quantify the violation of these condiations.

We compare:
- Loss reweighting: where a group (w_+) and/or a label (w_l) is weight is added to the loss;
- Coupled NNs strategy where one classifier per group is trained on the dataset and the ensemble is coupled using an L2 penalty.



The coupled NN strategy retains higher level of fairness without suffering severe loss of perfomance.
Preliminary tests on the CelebA benchmark confirmed the trend.



5. Bias mitigation

[1] Sarao Mannelli, Gerace, Rostamzadeh, Saglietti, arXiv:2205.15935.
[2] Suresh, Guttag, CoRR 2019.
[3] Ziwei, Luo, Wang, Tang, ICCV, 2015.
[4] Gerace, Saglietti, Sarao Mannelli, Saxe, Zdeborová, MLST, 2022.
[6] [follow-up work on the dynamics of bias] Jain, Nobahari, Baratin, Sarao Mannelli, arXiv:2405.18296.

6. References