UNFAIR GEOMETRIES: EXACTLY SOLVABLE DATA MODEL WITH FAIRNESS IMPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning (ML) may be oblivious to human bias but it is not immune to its perpetuation. Marginalisation and iniquitous group representation are often traceable in the very data used for training, and may be reflected or even enhanced by the learning models. In the present work, we aim at clarifying the role played by data geometry in the emergence of ML bias. We introduce an exactly solvable high-dimensional model of data imbalance, where parametric control over the many bias-inducing factors allows for an extensive exploration of the bias inheritance mechanism. Through the tools of statistical physics, we analytically characterise the typical properties of learning models trained in this synthetic framework and obtain exact predictions for the observables that are commonly employed for fairness assessment. Despite the simplicity of the data model, we retrace and unpack typical unfairness behaviour observed on real-world datasets. We also obtain a detailed analytical characterisation of a class of bias mitigation strategies. We first consider a basic loss-reweighing scheme, which allows for an implicit minimisation of different unfairness metrics, and quantify the incompatibilities between some existing fairness criteria. Then, we consider a novel mitigation strategy based on a matched inference approach, consisting in the introduction of coupled learning models. Our theoretical analysis of this approach shows that the coupled strategy can strike superior fairness-accuracy trade-offs.

1 INTRODUCTION

Machine Learning (ML) systems are actively being integrated in multiple aspects of our lives, from face recognition systems on our phones, to applications in the fashion industry, to high stake scenarios like healthcare. Together with the advantages of automatising these processes, however, we must also face the consequences of their — often hidden — failures. Recent studies Buolamwini & Gebru (2018); Weidinger et al. (2021) have shown that these systems may have significant disparity in failure rates across the multiple sub-populations targeted in the application. ML systems appear to perpetuate discriminatory behaviours that align with those present in our society Benjamin (2019); Noble (2018); Eubanks (2018); Broussard (2018). Discrimination over marginalised groups could originate at many levels in the ML pipeline, from the very problem definition, to data collection, to the training and deployment of the ML algorithm Suresh & Guttag (2021).

Data represents a critical source of bias Perez (2019). In some cases, the dataset can contain a record of a history of discriminatory behaviour, causing complex dependencies that are hardly eradicated even when the explicit discriminatory attribute is removed. In other cases (or even concurrently), the root of the discrimination can be found in the data collection process, and is related to the structural properties of the dataset. Heterogeneous representations of different sub-populations typically induce major bias in the ML predictions. Drug testing provides a historically significant example: substantial evidence Hughes (2007); Perez (2019) shows that the scarcity of data points corresponding to women individuals in drug-efficiency studies resulted in a larger number of side effects in their group.

In spite of a vast empirical literature, a large gap remains in the theoretical understanding of the bias-induction mechanism. A better theoretical grasp of this issue could help raise awareness and design more theoretically grounded and effective solutions. In this work, we aim to address this gap by introducing a novel synthetic data model, offering a controlled setting where data imbalances and the emergence of bias become more transparent and can be better understood.

To the best of our knowledge, the present study constitutes the first attempt to explore and exactly characterise by analytical means the complex phenomenology of ML fairness.

Summary of main results. We devise a novel synthetic model of data, the *Teacher-Mixture* (T-M), to obtain a theoretical analysis of the bias-induction mechanism. The geometrical properties of the model are motivated by common observations on the data structure in realistic datasets, concerning the coexistence of non-trivial correlations at the level of the inputs and between inputs and labels (some empirical observations can be found in appendix B). In particular, we focus on the role played by the presence of different sub-populations in the data, both from the point of view of the input distribution and from that of the labelling rule. Surprisingly, this simple structural feature is sufficient for producing a rich and realistic ML fairness phenomenology.

The parameters of the T-M can be tuned to emulate disparate learning regimes, allowing for an exploration of the impact of each bias-inducing factor and for an assessment of the effectiveness of a tractable class of mitigation strategies. In summary, in the present work we:

- Derive, through a statistical physics approach, an analytical characterisation of the typical performance of solutions of the T-M problem in the high-dimensional limit. The obtained learning curves are found to be in perfect agreement with numerical simulations in the same synthetic settings (as shown in the central panel in Fig. 1), and produce unfairness behaviours that are closely reminiscent of the results seen on real data.
- Isolate the different sources of bias (shown in the left panel of Fig. 1) and evaluate their interplay in the bias-induction mechanism. This analysis also allows us to highlight how unfairness can emerge in settings where the data distribution is apparently balanced.
- Trace a positive transfer effect between the different sub-populations, which implies that, despite their distinctions, an overall similarity can be exploited for achieving better performance on each group.
- Analyse the trade-offs between the different definitions of fairness, by studying the effects of a sample reweighing mitigation strategy, which can be encompassed in the theoretical framework proposed in this work and thus characterised analytically.
- Propose a model-matched mitigation strategy, where two coupled networks are simultaneously trained and can specialise on different sub-populations while mutually transferring useful information. We analytically characterise its effectiveness, finding that with this method, in the T-M, the competition between accuracy and different fairness metrics becomes negligible. Preliminary positive results are also reported on real data.

Further related works. In the past decade, algorithmic fairness has been receiving growing attention, spurred by the increasing number of ML applications in highly consequential social and economic areas Datta et al. (2015); Metz & Satariano (2020); Angwin et al. (2016). A central question in the field is on the proper mathematical definition of bias: the plethora of alternative fairness criteria includes measures of *group fairness*, e.g. statistical parity Corbett-Davies et al. (2017); Dwork et al. (2012); Kleinberg et al. (2016), disparate impact Calders & Verwer (2010); Feldman et al. (2015); Zafar et al. (2017b); Chouldechova (2017), equality of opportunity Hardt et al. (2016), calibration within groups Kleinberg et al. (2016), disparate mistreatment Zafar et al. (2017a), as well as measures of *individual fairness* Speicher et al. (2018); Castelnovo et al. (2022). We focus on group fairness in the following, since it is well-defined also in the high-dimensional limit considered in our theoretical framework. Recent works have highlighted incompatibilities between some of these fairness measures Kleinberg et al. (2016); Corbett-Davies & Goel (2018); Barocas et al. (2019), e.g. calibration and error disparity Pleiss et al. (2017), and their instability with respect to fluctuations in the training dataset Friedler et al. (2019); Castelnovo et al. (2022). Our work is the first to allow an exact quantification of the intrinsic trade-offs between these notions of group-fairness.

A second major topic in the field of algorithmic fairness is that of bias mitigation. In this work, we focus on *in-processing* strategies Arrieta et al. (2020), where the training process is altered in order to include fairness as a secondary optimisation objective for the learning model. These methods range from including *ad hoc* regularisation terms to the loss function Kamishima et al. (2012); Huang & Vishnoi (2019), to formulating fair classification as a constrained optimisation problem and deriving reduction-based algorithms Agarwal et al. (2018; 2019); Celis et al. (2019). Other possible strategies include adversarial training Zhang et al. (2018), where a fairness-arbiter model can drive learning towards a sough fairness criterion, and distributionally robust optimisation Słowik & Bottou (2021),

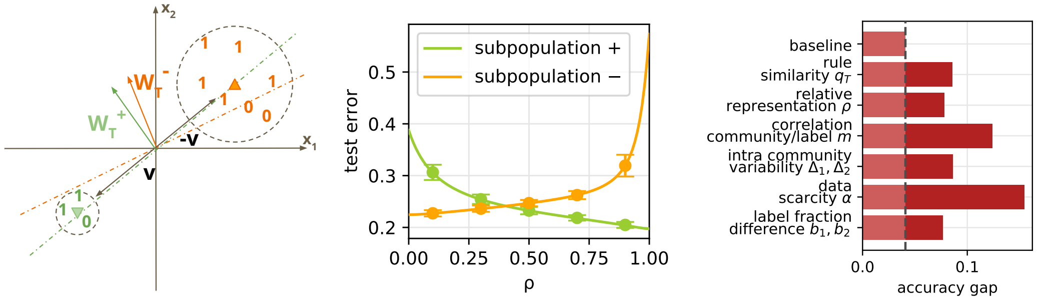

Figure 1: **T-M model**. Often we can distinguish sub-populations as clusters in a dataset according to some features. The label, e.g. effectiveness of a drug, is given by some rule acting on the data an may differ for the two subpopulations. In the T-M model, *(Left)*, the two sub-populations are drawn from two Gaussians around the two centres (green and orange triangles). The labels (plus and minus) are associated according to the hyper-planes $\boldsymbol{W}_T^+, \boldsymbol{W}_T^-$. In this 2D drawing we can see that the group $+$ (green) has 3 samples while group $-$ (orange) has 7 samples, so $\rho = 0.3$. The two hyper-planes are highly overlapping ($q_T \approx 1$) and weakly aligned with the shift vector ($m_+ \approx 0$, $m_- \approx 0$). Finally, we see that sub-population $+$ is less spread than sub-population $-$ ($\Delta_+ < \Delta_-$). *(Centre)* For this model, the test error can be calculated exactly, shown here as a function of ρ (solid curve). Numerical simulations (dots) closely match the analytical results. The panel exemplify the importance of ρ in creating bias against one sub-population or the other. *(Right)* Effect of changing one of the model parameter in terms of test accuracy gap, starting the from the set-up of the central panel with $\rho = 0.2$.

where one accounts for worst-case unfairness scenarios across the sub-populations in the data. In this work, we analyse two simple schemes whose performance can be analytically traced in our framework. First, an approach Kamiran & Calders (2012); Plecko & Meinshausen (2020); Lum & Johndrow (2016) based on loss-reweighing according to the associated subgroup and label of each data point. Second, we propose –and analyse– a novel method based on the introduction of coupled learning models, which can be interpreted as a modification of the ”two naive Bayes” model in Calders & Verwer (2010).

Alternative classes of debiasing approaches, which cannot be analysed within our framework, include pre-processing strategies Calmon et al. (2017); Feldman et al. (2015), learning unbiased representations Zemel et al. (2013), and post-processing techniques based on Decision Theory and Causal Reasoning Kamiran et al. (2012); Plecko & Meinshausen (2020).

2 MODELLING DATA IMBALANCE

The Teacher-Mixture model, sketched in Fig. 1, combines aspects of two common modelling frameworks for supervised learning, namely the Gaussian-Mixture (GM) and the Teacher-Student (TS) setups. The GM is a simple model of clustered input data, where each data point is sampled from one of a small set of –possibly overlapping– high-dimensional Gaussian distributions, while the TS provides a simple model of input-label correlation, where the ground-truth labels are obtained from a random ”teacher” neural network and the ”student” learning model tries to reproduce similar outputs. While retaining analytical tractability, the novel T-M data model allows for a richer phenomenology than the previous models, retracing the main features of real data with multiple coexisting sub-populations. For simplicity, the results discussed in this paper will focus on the case of two groups, but the analysis could be extended to multiple sub-populations.

More formally, we consider a synthetic dataset of n samples $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^n$, with $\mathbf{x}^\mu \in \mathbb{R}^d$, $y^\mu \in \{0, 1\}$. We define the $\mathcal{O}(1)$ ratio $\alpha = n/d$ and we refer to it as the data scarsity parameter. Each input vector is i.i.d. sampled from a mixture of two symmetric Gaussians with variances $\Delta = \{\Delta_+, \Delta_-\}$, $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{v}/\sqrt{d}, \Delta_\pm \mathbb{I}^{d\times d})$, with respective probabilities ρ and $(1-\rho)$. The shift vector $\mathbf{v}$ is a Gaussian vector with i.i.d. entries with zero mean and variance 1. The $1/\sqrt{d}$ scaling corresponds to the *high-noise* noise regime, where the two Gaussian clouds are overlapping and hard to disentangle Mignacco et al. (2020); Saglietti & Zdeborová (2022), e.g. as in the case of

CelebA and MEPS shown in appendix C. The ground-truth labels, instead, are provided by two i.i.d. Gaussian teacher vectors, namely $\mathbf{W}_T^+$ and $\mathbf{W}_T^-$, with components of zero mean and variance 1. Each teacher produces labels for the inputs with the corresponding group-membership, namely $y^\mu = \text{sign}\left(\mathbf{W}_T^\pm \cdot \mathbf{x}_\pm^\mu + \mathbf{b}_T^\pm\right)$. The thresholds $\mathbf{b}_T^\pm$ correspond to the teacher bias terms, included in the model to control the fraction of positive and negative samples within the two sub-populations. Overall, the geometric picture of the data distribution (a sketch in Fig. 1) is summarised by the following overlaps:

$$\tilde{m}_\pm = \frac{1}{d}\boldsymbol{W}_T^\pm \cdot \boldsymbol{v} \qquad q_T = \frac{1}{d}\boldsymbol{W}_T^+ \cdot \boldsymbol{W}_T^-, \tag{1}$$

that respectively quantify the alignment of the teacher decision boundaries with respect to the shift vector, controlling the group-label correlation, and the overlap between the teacher vectors, controlling the correlation between labels assigned to similar inputs belonging to different communities.

Given the synthetic dataset $\mathcal{D}$, we study the properties of a single-layer network trained via empirical risk minimization (ERM) of the loss:

$$\mathcal{L}(\boldsymbol{w}) = \sum_{\mu\in\mathcal{D}} \ell\left(\frac{\boldsymbol{T}_{c^\mu}\cdot\boldsymbol{x}^\mu}{\sqrt{d}} + \tilde{b}_{c^\mu}, \frac{\boldsymbol{w}\cdot\boldsymbol{x}^\mu}{\sqrt{d}} + b_s\right) + \frac{\lambda}{2}\left(\sum_{i=1}^{d} w_i^2\right) \tag{2}$$

where $\ell(y, \hat{y})$ is assumed to be convex, λ is an external parameter that regulates the intensity of the L_2 regularisation, and the index $c^\mu \in \{+, -\}$ denotes the group membership of data point μ. In this work, we derive a theoretical characterisation of the asymptotics of this learning model and consider the possible implications from a ML fairness perspective. In particular, we aim at studying the role of data geometry and cardinality in the training of a fair classifier.

Note that the T-M has, at the same time, the advantage of being simple, allowing better understanding of the many facets of ML bias, and the disadvantage of being simple, since some modelling assumptions might not reflect the complexity of real-world data. For example, we ignore any type of correlation among the inputs other than the clustering structure. The goal of this modelling work continues a long tradition of research in statistical physics, which has shown that theoretical insights gained in prototypical settings can often be helpful to disentangle and interpret the complexity of real world behaviour.

Remark 1 *By looking at the available degrees of freedom in the T-M, several possible sources of bias naturally emerge from the model:*

- *the* relative representation*, $\rho = n_+/(n_+ + n_-)$, with n_c the number of points in group c.*
- *the* group variance*, Δ_c, determining the width of the clusters.*
- *the* label frequencies*, controlled through the bias terms $\mathbf{b}_c$.*
- *the* group-label correlation*, m_c.*
- *the* labelling rule similarity*, q_T, which measures the alignment between the two teachers, i.e. the linear discriminators that assign the labels to the two groups of inputs.*
- *the* data scarcity*, α, representing the ratio between dataset size and input dimension.*

Theoretical analysis in high-dimensions. In principle, solving Eq. 2 requires finding the minimiser of a complex non-linear, high-dimensional, quenched random function. Fortunately, statistical physics Mézard et al. (1987) showed that in the limit $n, d \to \infty$, $n/d = \alpha$, a large class of problems, including the T-M model, becomes analytically tractable. In fact, in this proportional high-dimensional regime, the behaviour of the learning model becomes deterministic and trackable due to the strong concentration properties of a narrow set of descriptors that specify the relevant geometrical properties of the ERM estimator. The original high-dimensional learning problem can be reduced to a simple system of equations that depends on a set of scalar overlaps:

$$Q = \frac{1}{d}\boldsymbol{W}\cdot\boldsymbol{W}, \qquad m = \frac{1}{d}\boldsymbol{W}\cdot\boldsymbol{v} \qquad R_\pm = \frac{1}{d}\boldsymbol{W}\cdot\boldsymbol{W}_T^\pm, \tag{3}$$

representing the typical norm of the trained estimator, its magnetisation in the direction of the cluster centres, and its overlap with the two teachers of the T-M.

Analytical result 1 *In the high dimensional limit when* $n,\, d \to \infty$ *at a fixed ratio* $\alpha = n/d$, *the scalar descriptors* $\Theta = \{Q, m, R_\pm, \delta q\}$ *of the vector* $\boldsymbol{w}$ *obtained by the empirical risk minimisation of Eq. 2 with a convex loss, and their Lagrange multipliers* $\hat\Theta = \{\hat Q, \hat m, \hat R_\pm, \delta\hat q\}$, *converge to deterministic quantities given by the unique fixed point of the system:*

$$Q = -2\frac{\partial s(\hat\Theta;\lambda)}{\partial\,\delta\hat q};\quad m = \frac{\partial s(\hat\Theta;\lambda)}{\partial\,\hat m};\quad R_\pm = \frac{\partial s(\hat\Theta;\lambda)}{\partial\,\hat R_\pm};\quad \delta q = 2\frac{\partial s(\hat\Theta;\lambda)}{\partial\,\hat Q}; \tag{4}$$

$$\hat Q = 2\alpha\frac{\partial e(\Theta;\Delta)}{\partial\,\delta q};\quad \hat m = \alpha\frac{\partial e(\Theta;\Delta)}{\partial\,m};\quad \hat R_\pm = \alpha\frac{\partial e(\Theta;\Delta_\pm)}{\partial\,R_\pm};\quad \delta\hat q = 2\alpha\frac{\partial e(\Theta;\Delta)}{\partial\,Q}; \tag{5}$$

with:

$$s(\hat\Theta;\lambda) = \frac{\hat Q + \left(\hat m + \sum_{c=\pm}\tilde m_c \hat R_c\right)^2 + \sum_{c=\pm}(1-\tilde m_c^2)\hat R_c^2 + 2\left(q_T - \prod_{c=\pm}\tilde m_c\right)\prod_{c=\pm}\hat R_c}{2\left(\delta\hat q + \lambda\right)} \tag{6}$$

$$e(\Theta;\Delta) = \mathbb{E}_c\left[\mathbb{E}_z\sum_{y=\pm1} H\left(-y\frac{\sqrt{Q}(c\,\tilde m_c + \tilde b_c) + \sqrt{\Delta_c}R_c z}{\sqrt{\Delta_c(Q-R_c^2)}}\right) v(y,c,\Theta)\right] \tag{7}$$

where $c \in \{+,-\} \sim \mathrm{Bernoulli}(\rho)$, $z \sim \mathcal{N}(0,1)$, $H(\cdot) = \frac{1}{2}\mathrm{erfc}(\cdot/\sqrt{2})$ *is the Gaussian tail function,* w *is the solution of:*

$$v(y,c,\Theta) = \max_w\left[-\frac{w^2}{2} - \ell\left(y, \sqrt{\Delta_c\delta q}w + \sqrt{\Delta_c Q}z + c\,m + b\right)\right] \tag{8}$$

and the bias b *implicitly solves the equation* $\partial_b e(\Theta;\Delta) = 0$.

Note that this result was obtained through the non rigorous yet exact replica method from statistical physics Mézard et al. (1987); Engel & Van den Broeck (2001); Zdeborová & Krzakala (2016). The derivation details are deferred to appendix D. We remark that several analytic results obtained through the replica method have been subsequently proved rigorously. In particular, the proofs presented by Thrampoulidis et al. (2015); Mignacco et al. (2020); Loureiro et al. (2021) in settings similar to the present one suggest that an extension for the T-M case could be derived. However, this is left for future work. In this manuscript, we verify the validity of our theory by comparison with numerical simulations, as shown e.g. in the central panel of Fig. 1.

The obtained fixed point for the scalar descriptors Θ can be used to evaluate simple expressions for common model evaluation metrics, such as the *confusion matrix* or the *generalisation error*.

Analytical result 2 *In the same limit as in Analytical result 1, the entries of the confusion matrix, representing the probability of classifying as* $\hat y$ *an instance sampled from sub-population* c *with true label* y, *are given by:*

$$p\left(\hat y \,|\, y; c\right) = \mathbb{E}_z\left[\mathrm{Heav}\left(y\left(\sqrt{\Delta_c}z + c\,\tilde m_c + \tilde b_c\right)\right) H\left(-\hat y\frac{(c\,m + b) + \sqrt{\Delta_c}R_c z}{\sqrt{\Delta_c(Q-R_c^2)}}\right)\right], \tag{9}$$

where $z \sim \mathcal{N}(0,1)$ *and* $\mathrm{Heav}(\cdot)$ *is the Heaviside step function. The generalization error, representing the fraction of wrongly labelled instances, can then be obtained as* $\epsilon_g = \mathbb{E}_c\left[\sum_{\hat y \neq y} p(\hat y \,|\, y; c)\right]$.

This second result provides us with a fully deterministic estimate of the accuracy of the trained model on the different data sub-populations. These scores will be used in the following sections to investigate the possible presence of bias in the classification output of the model. Note that theorems 1 and 2 allow for an extremely efficient and exact evaluation of the learning performance in the T-M, remapping the original high-dimensional optimisation problem onto a system of deterministic scalar equations that can be easily solved by recursion.

3 INVESTIGATING THE SOURCES OF BIAS

With these analytical results at hand, we now turn to systematically investigating the effect of the sources of bias identified in remark 1, which potentially mine the design of a fair classifier. To

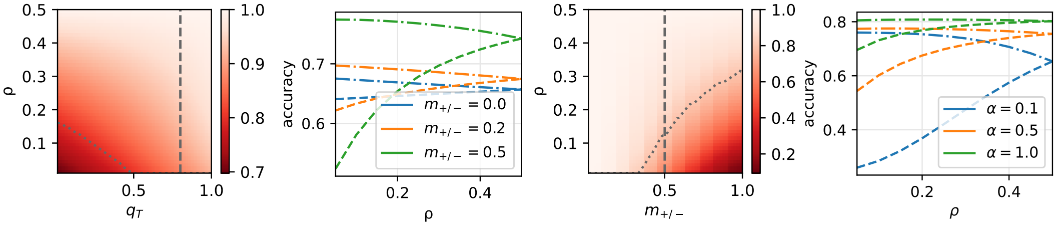

Figure 2: **Bias under different parametric settings.** Impact of several parameters on the Disparate Impact (DI) of the model. From left to right. *(Panel 1)* Phase diagram where each point represents the DI (red indicates a worse accuracy on group $+$) for different values of rule similarity q_T (x-axis) and the relative representation ρ (y-axis). The dotted grey line denotes the 80% threshold for disparate impact. *(Panel 2)* Accuracy of group $+$ (dashed lines) and group $-$ (dot-dashed lines), in a cut across the first phase diagram at $q_T = 0.8$. The different colours indicate different levels of group-label correlation $m_\pm$. *(Panel 3)* Phase diagram of the DI at fixed $q_T = 1$, as the group-label correlation $m_\pm$ (x-axis) and ρ are varied. *(Panel 4)* Role of the dataset size (α), at a cut $m_\pm = 0.5$ of the diagram in panel 3.

quantify the level of bias in the predictions of the trained model, we need to choose a metric of fairness. Throughout this section, we employ *disparate impact* (DI) Feldman et al. (2015), a ML analogous of the 80% rule Commission et al. (1979), which allows a simple assessment of the over-specialisation of the classifier on one of the sub-populations. In principle, in the T-M there is no preferable realisation of the target attribute so we can adopt a symmetric version of DI, defined as the ratio between test accuracy in sub-population $+$ and sub-population $-$. We consider three separate experiments to summarise some distinctive features of the fairness behaviour in the T-M: namely, the impact of the correlation between the labelling rules and the group structure, the interplay between relative representation and group variance, and the positive transfer effect in the data-scarse regime. The parameters of the experiments, if not specified in the caption, are detailed in appendix E.1.

Group-label correlation. In the two left panels of Fig. 2, we consider a scenario where the labelling rules for the two groups are not perfectly aligned, i.e. $\boldsymbol{W}_T^+ \neq \boldsymbol{W}_T^-$ (and/or $b_+ \neq b_-$). Note that in this case we have a clear mismatch between the learning model, a single linear classifier, and the true input-output structure in the data: the learning model cannot reach perfect generalisation for both sub-populations at the same time. For simplicity, we set an equal correlation between the two teacher vectors and the shift vector, $m_+ = m_- > 0$, and isolate the role of rule similarity q_T. The first panel shows a phase diagram of the DI (DI<1 indicating a lower accuracy on group $+$), as function of the similarity of the teachers and the fraction of $+$ samples in the dataset. As intuitively expected, the induced bias exceed the 80% rule when the labelling rules are misaligned and the group sizes are numerically unbalanced (small q_T and ρ). Indeed, in the cut displayed in the second panel, by lowering the group-label correlation $m_\pm$ the gap between the measured accuracies on the two sub-populations becomes smaller. However:

Remark 2 *Even when $q_T = 1$ and the task is solvable (i.e. the classifier can learn the input-output mapping), the trained model can still be biased.*

This is shown in the two panels on the right of Fig. 2, where a large high-bias region (DI$<80\%$) exists. In particular, the third panel shows the cause of this effect in the presence of a non-zero group-label correlation $m_\pm$, and in the fourth panel we see how this effect is more pronounced in the data-scarse regime. In all four panels, as ρ reaches 0.5, the two sub-populations become equally represented and the classifier achieves the same accuracy for both.

Bias and variance. In Fig. 3, we plot the DI as a function of the group variances $\Delta_\pm$, for different values of the fraction of $+$ samples. One finds that the model might need a disproportionate number of samples in the two groups to obtain comparable accuracies. We can see that:

Remark 3 *Balancing the group relative representation does not guarantee a fair training outcome.*

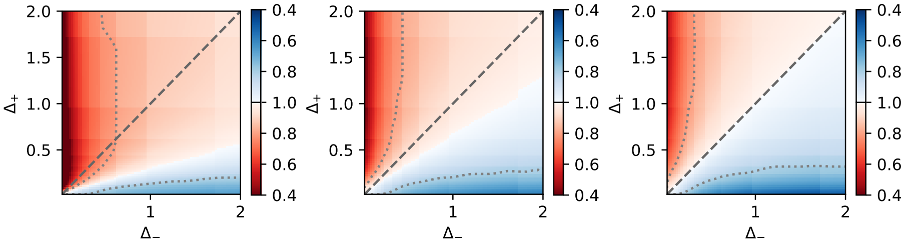

Figure 3: **Bias in equally represented subpopulations.** We show the disparate impact as the distribution of the two subpopulations is changed by altering their variances (Δ_+ and Δ_-). The diagonal line gives the configurations where the two subpopulations have the same variance. The three panels consider different levels of representation, from left to right $\rho = 0.1, 0.3, 0.5$. The latter is the situation with both subpopulations being equally represented in the dataset. We use the red and blue colours to quantify the disparate bias against sub-population $+$ and $-$ (respectively).

In fact, the quality of a group's representation in the dataset can increase if the number of points is kept constant but the group variance is reduced. The blue regions in the first two panels indicate a higher accuracy for the minority group even if the dataset only contains 10% and 30% of samples belonging to it. This exemplifies the fact that a very focused distribution (low $\Delta_\pm$) actually requires less samples. The last panel ($\rho = 0.5$) shows the scenario one would expect *a priori*: on the diagonal line the DI is balanced, but by setting $\Delta_+ > \Delta_-$ (or viceversa) one induces a bias in the classification.

Positive transfer. If mixing different sub-populations in the same dataset can induce unfair behaviour, one could think of splitting the data and train independent models. In Fig. 4, we show that a *positive transfer effect* Gerace et al. (2022) can yet be traced between the two groups when the rules are sufficiently similar.

Remark 4 *The performance on the smaller group tends to further deteriorate if the dataset is split according to the sub-group structure.*

To clarify this point, we plot the DI as a function the data scarcity α, for several values of the rule similarity q_T and at fixed ρ. We also compare the accuracies on each sub-population of a classifier trained on the full dataset and of a baseline classifier trained only on the respective data subsets ($+$ in the second figure, $-$ in the third). If the rules are sufficiently similar (large q_T), we can observe a positive transfer and using the dataset in his entirety leads to a performance and fairness improvement. As expected, positive transfer can be particularly useful in data-scarce regimes (small α) and becomes ineffective or detrimental in large datasets (large α), as shown in the last panel.

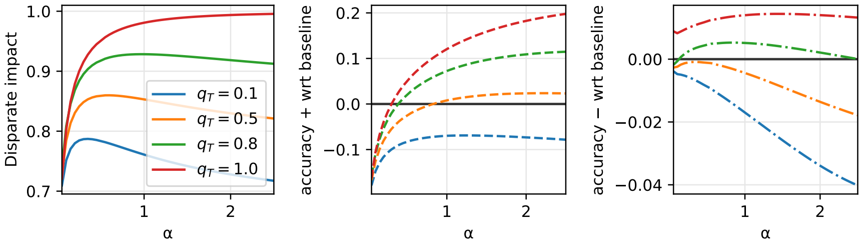

Figure 4: **Positive transfer effect.** Given a fixed proportion of the two sub-populations, we compare different levels of rule similarity (q_T) as the size of the dataset is increased. The accuracy gap (first figure) may mislead into thinking that the accuracy in one sub-population is decreasing as the other increases, instead the accuracy is steadily increasing (second figure) for both sub-populations. Finally, the last two figures show the accuracy in the sub-population $+$ and $-$ (respectively) minus the accuracy on the same dataset when the other sub-population is perfectly removed.

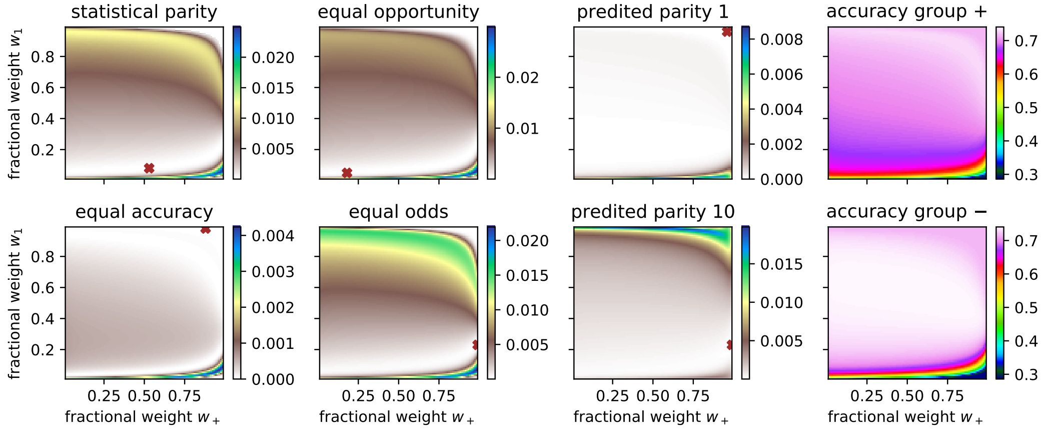

Figure 5: **Mitigation trade-off.** The phase diagrams show the effect of re-weighting, biasing both: towards low mistakes in classifying sub-population $+$ (w_+ on the x-axis) and towards low mistakes for label $+1$ (w_1 on the y-axis). The quantity shown in the diagrams are the mutual information for the metrics introduced in the text (first three columns) and the accuracy on the two sub-populations (last column). The diagrams for the mutual information also show red markers denoting where the minimum is achieved.

4 MITIGATION STRATEGIES

To assess the fairness of a ML model on a given data distribution, a plethora of different fairness criteria have been designed Speicher et al. (2018); Castelnovo et al. (2022). Appendix F presents a summary of the criteria considered in our analysis. Following the lines of Speicher et al. (2018), we aim to quantify exactly how far is a given trained model from meeting each of these criteria. Given a classification event E –specified by the criterion– and the group membership C, a natural measure of their independence is provided by the Mutual Information (MI):

$$I(E;C) = D_{KL}(\mathbb{P}[E,C] \mid \mathbb{P}[E]\mathbb{P}[C]) = \mathbb{E}_{(E,C)} \log \frac{\mathbb{P}[E,C]}{\mathbb{P}[E]\mathbb{P}[C]}. \tag{10}$$

Clearly, the fairness condition is completely verified only if the joint distribution factorises, i.e. $\mathbb{P}[E,C] = \mathbb{P}[E]\mathbb{P}[C]$, and the mutual information goes to zero. This represents the impossibility of predicting the classification outcome of an unbiased model just from the group membership.

In the following, we consider two simple bias mitigation strategies that can be analysed within our analytical framework. The required generalisations of the replica results are detailed in appendix D. First, we study the de-biasing effect of a sample reweighing strategy where the relevance of each sample is varied based on its label and group membership Kamiran & Calders (2012); Plecko & Meinshausen (2020); Lum & Johndrow (2016). By adjusting the weights, one can indirectly minimise the MI relative to any given fairness measure. We use the simultaneous quantitative predictions on the various metrics to assess the compatibility between different fairness definitions. Then, we propose a theory-based mitigation protocol, along the lines of protocols used in the context of multi-task learning Rusu et al. (2016).

Loss Reweighing. Recent literature shows that some fairness constraints cannot be satisfied simultaneously. ML systems are instead forced to accept trade-offs between them Kleinberg et al. (2016). This sort of compromise is well-captured in the simple framework of the T-M model. Fig. 5 shows, in form of phase diagrams, the MI measured with respect to the various fairness criteria while varying the two reweighing parameters, w_1 and w_+, which up-weigh data points with true label 1 and in group $+$, respectively. E.g., the loss term associated to a label 1-group $+$ sample will be weighed $w_+ w_1$, while that of a label 0-group $-$ data point will receive weight $(1 - w_+)(1 - w_1)$. By changing these relative weights one can force the model to pay more attention to some types of errors and re-establish a balance between the accuracies on the two sub-populations. The red crosses in the phase diagrams identify the points where the MI reaches its minimum value for each fairness metric. Notably, some minima are found to lie in different regions of the phase diagram (at the opposite extremes), and they often align only in correspondence of trivial classification, where

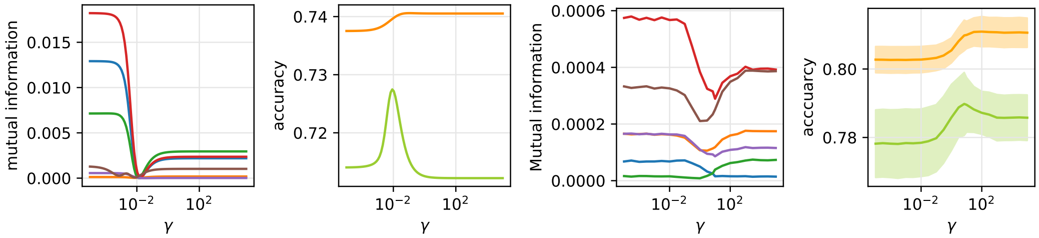

Figure 6: **Mitigation trade-off in the coupled architecture.** The first two figures represent a one dimensional version of Fig. 5 for the coupled architecture set up. On the left panel, the mutual information of the different fairness measures (statistical parity, equal opportunities, equal accuracy, equal odds, predicted parity 1, predicted parity ± 1) is plot as function of the coupling-strength parameter γ, observe that the minima of the curves are much closer. Furthermore, the second panel shows a better accuracy trade-off between subpopulation $+$ and subpopulation $-$. The remaining two figures, show an example from the CelebA dataset splitting and classifying according to the attributs ”Wearing_Lipstick” and ”Wavy_Hair” respectively, more details are provided in appendix B and C. The observations made for the synthetic model applies also in this real-world case.

fairness is achieved but at the expense of accuracy. These results are in agreement with rigorous results in the literature Barocas et al. (2019), but also show how the incompatibilities between the different constraints extend to regimes where the fairness criteria are not exactly satisfied.

Coupled Networks. The emergence of classification bias in the T-M could be lead back to the clear mismatch between the generative model of data and the learning model. In order to move towards a matched inference setting, we need to enhance the learning model to account for the presence of multiple sub-populations and labelling rules. This inspires a novel mitigation strategy – called *coupled neural networks*. The strategy consists in the simultaneous training of multiple neural networks, each one seeing a different subset of the data associated with a different sub-population. The networks exchange information by means of an elastic penalty that mutually attracts them, and the intensity of this elastic interaction is obtained by cross-validation. This approach is close in spirit to other methods already present in the literature Calders & Verwer (2010); Saglietti et al. (2021); Zenke et al. (2017).

Remark 5 *The coupled neural networks method allows for higher expressivity and specialisation on the various sub-populations, while also encouraging a positive transfer between similarly labelled sub-groups, leading to better fairness-accuracy trade-offs*

The first plot in Fig. 6, displaying the behaviour of the mutual information as a function of the coupling parameter for different fairness metrics, shows the key advantage of using this method. We observe is a more robust consistency among the various fairness metrics: the positions of the different minima are now very close to each other. Moreover, the value of the coupling parameter achieving this agreement condition is also the one that minimises the gap in terms of test accuracy between the two sub-populations, as shown in the second plot of Fig. 6, without hindering the performance on the larger group. Notice that this result does not contradict the impossibility theorem Barocas et al. (2019) which states that statistical parity, equal odds, and predicted parity cannot be satisfied altogether. In fact, our result only concerns soft minimisation of each fairness metrics. In appendices D and F.1 we provide additional results for this method and we discuss the effect of training the networks on data subsets that only partially correlate with the true group structure.

Despite the fact that the T-M is just a data prototype, the positive agreement with real phenomenology suggests that this method could be effective also on real-world data. The remaining two plots in Fig. 6 show preliminary results of the performance of the coupled neural networks strategy in the realistic dataset from CelebA[1]. We stress that although the method works significantly better in the synthetic framework, real data present more complex correlations that may hinder the effectiveness of the method. Therefore, an application of this technique on real settings requires caution. A future research direction will be to understand the range of applicability of the coupled neural networks and, consequently, its limitations.

---

[1] The illustrated chekpoints are used only to show the similarity of behavior in synthetic data and realistic data (CelebA), and not used or recommended to use in any face recognition systems or scenarios.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pp. 120–129. PMLR, 2019.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social Forces*, 2019.

Lynn A. Blewett, Julia A. Rivera Drew, Risa Griffin, Natalie Del Ponte, and Pat Convey. IPUMS health surveys: Medical expenditure panel survey, version 2.1 [dataset]. *Minneapolis, MN: IPUMS*, 2021. URL `https://doi.org/10.18128/D071.V2.1`.

Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world*. mit Press, 2018.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21, 2022.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

US Equal Employment Opportunity Commission et al. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *US Equal Employment Opportunity Commission: Washington, DC, USA*, 1979.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015. doi: doi: 10.1515/popets-2015-0007. URL `https://doi.org/10.1515/popets-2015-0007`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pp. 2879–2890. PMLR, 2019.

Robert N Hughes. Sex does matter: comments on the prevalence of male-only investigations of drug effects on rodent behaviour. *Behavioural pharmacology*, 18(7):583–589, 2007.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. *arXiv preprint arXiv:2205.05256*, 2022.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 35–50. Springer, 2012.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

Cade Metz and Adam Satariano. An algorithm that grants freedom, or takes it away. *The New York Times*, 6, 2020.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pp. 6874–6883. PMLR, 2020.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, pp. 3, 2018.

Safiya Umoja Noble. *Algorithms of oppression*. New York University Press, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Caroline Criado Perez. *Invisible women: Data bias in a world designed for men*. Abrams, 2019.

Drago Plecko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Negar Rostamzadeh, Ben Hutchinson, Christina Greer, and Vinodkumar Prabhakaran. Thinking beyond distributions in testing machine learned models. *arXiv preprint arXiv:2112.03057*, 2021.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Luca Saglietti and Lenka Zdeborová. Solvable model for inheriting the regularization through knowledge distillation. In *Mathematical and Scientific Machine Learning*, pp. 809–846. PMLR, 2022.

Luca Saglietti, Stefano Sarao Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher-student networks. *arXiv preprint arXiv:2106.08068*, 2021.

Agnieszka Słowik and Léon Bottou. Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation. *arXiv preprint arXiv:2106.09467*, 2021.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248, 2018.

Harini Suresh and John Guttag. Understanding potential sources of harm throughout the machine learning life cycle. 2021.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

CONTENTS

A EXTENDED BROADER IMPACT STATEMENT

According to the Cambridge dictionary, bias is "the action of supporting or opposing a particular person or thing in an unfair way". Bias is inherently rooted in societies, and each society carries their prejudice in favour and against some groups compared with others. We acknowledge that defining or addressing biases in a social system only from a technical standpoint, may risk the root causes of issues and even amplify other types of biases that are dismissed. We also recognise there are multiple incompatible notions of Fairness and Bias Narayanan (2018). Thus, we recognise it is very important to define and identify bias and fairness definitions that fit the task and context in hand. With this paper, our aim is to quantify biases that stem from data imbalance in a system in a controlled setting. This will help identify and mitigate biases that stem from geometric properties of the input data or the insufficient number of quality labels for some subgroups.

It is important to recognise the priority of debiasing in a system and perform appropriate bias tests considering the context and task in hand Rostamzadeh et al. (2021); Hutchinson et al. (2022). In our work, this indicates the clusters in represented data space being associated with attributes that matter in the system for the task of interest. In a software system, unit testing is performed to identify edge cases. However, in ML systems this search space might be larger and identifying the features that we want to debias the systems with, is of importance.

On the choice of the datasets, in this paper, we used synthetic data to have control over the experiments and hypothesis. We also used CelebA dataset to compare a realistic data representation with our synthetic set up. We acknowledge that CelebA datasets have features like gender, and age that may not be representative or inclusive. For example the binary gender attributes may be harmful to trans and gender non-conforming communities. Therefore, we only used features like "Wavy_Hair" and "Wearing_Lipstick" that don't inherit social constructs and are appearance based. We also state that the outcome of this research should not be used in any ways or forms in a face recognition or detection system.

B MODEL MOTIVATION

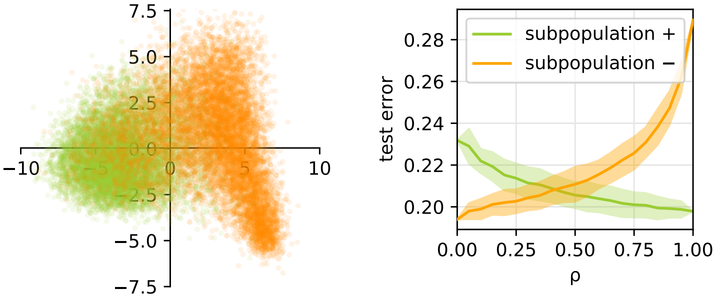

Figure B.1: **Relative representation and bias**. Numerical experiments on a sub-sample of the CelebA dataset. *(Left)* A 2D projection of the pre-processed dataset, obtained from PCA, where the colours represent the two sub-populations. *(Right)* Per community test error, as the fraction of samples from the two subpopulations is varied (dataset dimension is fixed).

Consider the following simple experiment. The CelebA dataset Liu et al. (2015) is a collection of face images of celebrities, equipped with metadata indicating the presence of specific attributes in each picture. We construct a dataset by sub-sampling CelebA and by preprocessing the selected images through an Xception network Chollet (2017) trained on ImageNet Deng et al. (2009). As depicted in the scatter plot in Fig. B.1, the first two principal components of the obtained data clearly reveal a clustered structure. Many attributes contained in the metadata are highly correlated with the split into these two sub-populations. For example, in the figure we colour the points according to the attribute ”Wearing_Lipstick”. Now, suppose we are interested in predicting a different target attribute, which is not as easily determined by just looking at the group membership, e.g. ”Wavy_Hair”[2]. What happens to the model accuracy if one alters the *relative representation* of the two groups, e.g. when one varies the fraction of points that belong to the orange group?

The right panel of Fig. B.1 shows the outcome of this experiment. As we can see from the plot, the fact that a group is under-represented induces a gap in the generalisation performance of the model when evaluated on the different sub-populations. The presence of a gap is a clear indicator of unfairness, induced by an implicit bias towards the over-represented group.

Many factors might play a role in determining and exacerbating this phenomenon. This is precisely why designing a general recipe for a fair / unbiased classifier is a very challenging, if solvable, problem. Some bias inducing factors are linked to the sampling quality of the dataset, as in the case of the overall number of datapoints and the balance between the sub-populations frequencies. Other factors are controlled by the different degree of variability in the input distributions of each group. In other cases the imbalance is hidden and can only be recognised by looking at the joint distribution of inputs and labels. For example, the balance between the positive/negative labels might differ among the groups and may be strongly correlated with the group membership. Even similar individuals with different group memberships might be labelled differently. The present work aims at modelling the data structure observed in these types of experiments, to obtain detailed understanding of the various sources of bias in these problems.

---

[2] To be mindful on the Ethical Considerations of using the CelebA datast, we don’t use protected attributes like binary genders and age Denton et al. (2019)

C REAL DATA VALIDATION

In this section, we provide extra details concerning the experiment conducted on the CelebFaces Attributes (CelebA) dataset and described in the main text. We also show how the phenomena presented in the main text are quite general and can be observed even in lower-dimensional datasets, such as the Medical Expenditure Panel Survey (MEPS) dataset Blewett et al. (2021).

C.1 ADDITIONAL DETAILS ON THE CELEBA EXPERIMENTS

The CelebA dataset is a collection of 202.599 face images of various celebrities, accompanied by 40 binary attributes per image (for instance, whether a celebrity features black hairs or not) Liu et al. (2015). To obtain the results presented in the main text we apply the following pre-processing pipeline: We first downsample CelebA up to 20.000 images. Notice that this is done with the purpose of considering settings with limited amount of available data. Indeed, as we have seen in the main manuscript, data scarcity is one of the main bias-inducing ingredients. We are thus not interested to consider the entire CelebA dataset, especially for simple classification tasks like the one described in the main text. By exploiting the deep learning framework provided by Tensorflow Abadi et al. (2015), we then pre-process the dataset using the features extracted from an Xception convolutional network Chollet (2017) pre-trained on Imagenet Deng et al. (2009). Finally, we collect the extracted features together with the associated binary attributes in a json file.

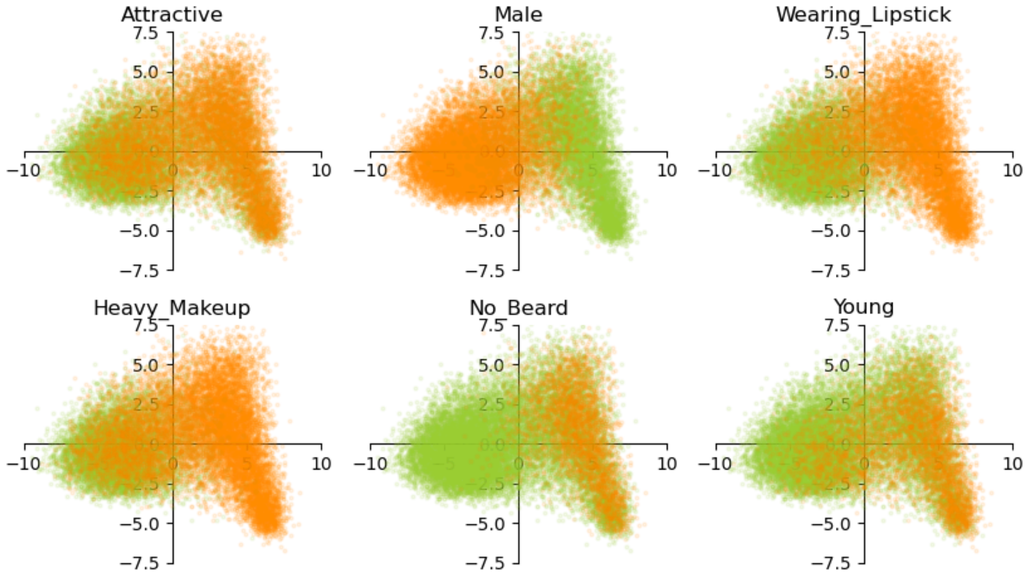

Figure C.1: **Clustering CelebA according to attributes.** We show 6 of the 40 attributes in CelebA demonstrating a neat clustering.

By applying PCA on the pre-processed dataset, we observe a clustering structure in the data when projected to the space of the PCA principal components. The clusters appear to reflect a natural correspondence with the binary attributes associated to each input data point, however this is not a general implication and many datasets show clustering with a non interpretable connection to the attributes. The clusters can be clearly seen in Fig. C.1, where we use colours to show whether a celebrity features a given attribute (green dots) or not (orange dots). In the plot, the axes correspond to the directions traced by the two PCA leading eigenvectors. As we can see from Fig. C.1, the two subpopulations are overlapping and hard to disentangle. This situation precisely corresponds to the high-noise regime the T-M model is meant to describe. Among the various clustering depicted in Fig. C.1, we decided to disregard those corresponding to ethically questionable attributes, such as ”Attractive”, ”Male” or ”Young”. Finally, we chose as sensitive attribute – determining the membership in the subpopulations – the ”Wearing_Lipstick” feature since it gives a more homogeneous distribution of the data points in the two clusters.

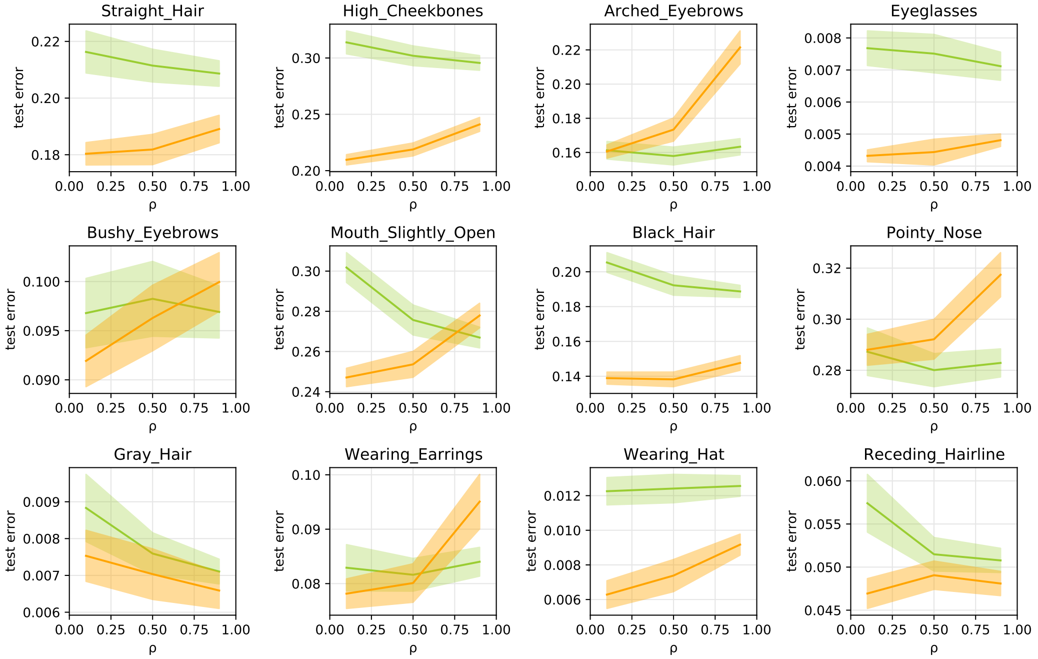

Figure C.2: **Relative representation across attributes.** The panels show the generalisation error depending on the relative representation in different attributes. The sub-populations $+$ (green) $-$ (orange) are obtained splitting according to the attribute ”Wearing_Lipstick”. The simulations are averaged over 100 samples.

Anyone of the other attributes can be considered as a possible target, and thus be used to label the data points. The final pre-processing step consist in downsampling further the data in order to have the same ratio of 0 and 1 labels in the two subpopulations. This step helps mitigating bias induced by the different ratio of label in the two subpopulations and simplifies the identification of the other sources of bias. The general case can be addressed in the T-M model, in Sec. E we comment more on the bias induced by different label ratios.

As Fig. C.2 illustrates, there is a large number of possible outcomes concerning the behaviour of the test error as a function of the relative representation. Indeed, as we have seen in the main text, the presence and the position of the crossing point strictly depends on both the cluster variances and the amount of available data. Despite all these behaviours are fully reproducible in the T-M model by means of its corresponding parameters, we here decided to chose the “Wavy_Hair” as target feature because it shows a nicely symmetric profile of the test error that is more suitable for illustration purposes. To get the learning curves in Fig. C.2, we train a classifier with logistic regression and L_2-regularization. In particular, we use the LogisticRegression class from scikit-learn Pedregosa et al. (2011). This class implements several logistic regression solvers, among which the *lbfgs* optimizer. This solver implements a second order gradient descent optimization which can consistently speed-up the training process. The training algorithm stops either if the maximum component of the gradient goes below a certain threshold, or if a maximum number of iterations is reached. In our case, we set the threshold at $1e-15$ and the maximum number of iterations to 10^5. The parameter *penalty* of the LogisticRegression class is a flag determining whether an L_2-regularization needs to be added to the training or not. The C hyper-parameter corresponds instead to the inverse of the regularization strength. In our experiments, we chose the value of the regularization strength by cross-validation in the interval $(10^{-3}, 10^3)$ with 30 points sampled in logarithmic scale.

C.2 OTHER DATASETS

The observations made on the CelebA dataset are quite general and can be further extended to lower-dimensional datasets. To demonstrate this, we considered the Medical Expenditure Panel Survey

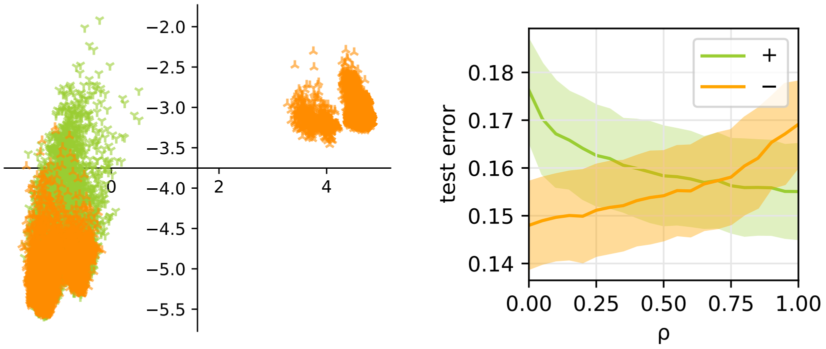

Figure C.3: **MEPS dataset.** (Left) Clustering in the MEPS dataset, according to be above or below the average age. (Right) Crossing of the generalisation error as the relative representation ρ is changed. The simulations are averaged over 100 samples.

(MEPS) dataset. This is a dataset containing a large set of surveys which have been conducted across the United States in order to quantify the cost and use of health care and health insurance coverage. The dataset consists of about 150 features, including sensitive attributes, such as age or medical sex, as well as attributes describing the clinical status of each patient. The label is instead binary and measures the expenditure on medical services of each individual, assessing whether the total amount of medical expenses is below or above a certain threshold. As it can be seen in Fig. C.3, the behaviour is qualitatively similar to the one already observed in the CelebA dataset of celebrity face images. Indeed, even in this case, PCA shows the presence of two distinct clusters when considering the age as the sensitive attribute and then splitting the dataset in two sub-populations, according to the middle point of the age distribution. Moreover, the generalisation error per community exhibits a crossing according to the relative representation.

D REPLICA APPROACH AND ANALYSIS

DATA MODEL

We consider a classification problem defined by a set of input-output associations $\mathcal{D} = \{(\mathbf{x}^\mu, y^\mu)\}_{\mu=1}^{\alpha N}$, with $\mathbf{x}^\mu \in \mathbb{R}^N$, $y^\mu \in \{0, 1\}$ and $\alpha = \mathcal{O}(1)$. Note that here we are employing the statistical physics notation, indicating the number of features (and the number of trained parameters) with N, whereas in the main text we used the statistics notation d.

The data is generated according to the Teacher-Mixture (T-M) model described in the main text. The inputs $\mathbf{x}^\mu$ are distributed as in a Gaussian mixture: first, one samples the group-membership according to:

$$c^\mu \sim \rho\, \delta\left(c^\mu - 1\right) + (1-\rho)\, \delta\left(c^\mu + 1\right) \tag{D.1}$$

where $0 < \rho < 1$ is the fraction of samples from the first Gaussian. Then, the inputs are generated as:

$$\boldsymbol{x}^\mu = c^\mu \frac{\boldsymbol{v}}{\sqrt{N}} + \boldsymbol{z}^\mu \tag{D.2}$$

where the shift-vector $\mathbf{v}$ and the noise-vector $\mathbf{z}^\mu$ have i.i.d. components $v_i \sim \mathcal{N}\left(0,1\right), z_i \sim \mathcal{N}\left(0, \Delta_{c^\mu}\right)$, with a variance Δ_{c^μ} that is group dependent. The $1/\sqrt{N}$ scaling for the shift of the centres is the interesting one where the Gaussians might have a large overlap.

The ground-truth labels y^μ are instead assigned by a tuple of teacher vectors, each acting on the patterns in the relative community:

$$y^\mu = \mathrm{sign}\left(\frac{\mathrm{T}_{\mathrm{c}^\mu} \cdot \mathrm{x}^\mu}{\sqrt{\mathrm{N}}} + \tilde{\mathrm{b}}_{\mathrm{c}^\mu}\right) \tag{D.3}$$

where $T_i \sim \mathcal{N}(0,1)$ and $\tilde{b}_{c^\mu}$ are bias terms of order $\mathcal{O}(1)$. Note that the $1/\sqrt{N}$ scaling ensures that the activation is $\mathcal{O}(1)$ overall. The teacher vectors are assumed to have a fixed overlap with the centres direction, $\tilde{m}_c = \frac{\boldsymbol{T}_c \cdot \boldsymbol{v}}{N}$, and mutual overlap $\tilde{q} = \frac{\mathbf{T} \cdot \mathbf{T}}{N}$. Since the geometry of the problem is completely determined by these quantities, in the following we will call them the generative parameters.

LEARNING MODEL

We will directly present the most general setting for this calculation, where the learning model is composed of two linear classifiers, coupled by an elastic penalty of intensity γ. This allows us to characterise the novel mitigation strategy proposed in this work, while the standard case with a single learning model can be obtained by setting $\gamma = 0$. The derivation presents elements of novelty that are interesting *per se* and we will publish a more technical version of the paper to highlight these new results.

Each student, denoted by the index $s = 1, 2$, is trained only on a fraction of the full dataset $\mathcal{D}_s$, and obtains information on the rest through the coupling with the second learning model. Note that the data split is not assumed to be perfectly aligned with the group structure, despite our intuition that this might allow the best generalisation performance.

The loss function for the coupled learning model reads:

$$\mathcal{L}(\boldsymbol{w}_1, \boldsymbol{w}_2) = \sum_{s=1,2} \sum_{\mu \in \mathcal{D}_s} \ell\left(\frac{\boldsymbol{T}_{c^\mu} \cdot \boldsymbol{x}^\mu}{\sqrt{N}} + \tilde{b}_{c^\mu}, \frac{\boldsymbol{w}_s \cdot \boldsymbol{x}^\mu}{\sqrt{N}} + b_s\right) + \sum_{s=1,2} \frac{\lambda}{2}\left(\sum_{i=1}^{N} w_{s,i}^2\right) - \frac{\gamma}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 \tag{D.4}$$

Specifically in the following we will focus on the cross-entropy loss:

$$\ell(y, q) = -\Theta\left(y\right) \log \sigma(q) - \left(1 - \Theta\left(y\right)\right) \log\left(1 - \sigma(q)\right) \tag{D.5}$$

where $\Theta(\cdot)$ is the Heaviside step function, which outputs 1 for positive arguments and 0 for negative ones, and $\sigma\left(x\right) = \left(1 + \exp\left(-x\right)\right)^{-1}$ is the sigmoid activation function. The calculation holds also for alternative losses, e.g. the Hinge loss or the MSE loss, since the only affected part is the numerical optimisation of the proximal operator, as we show below.

AVERAGE OVER THE MEASURE OF COUPLED TEACHERS

In the T-M model, the label distribution is non-trivially dependent on the mutual alignment of the shift vector $\boldsymbol{v}$, determining the means of the two Gaussians in the input mixture, and the two teacher vectors $\boldsymbol{T}_{1,2}$. As in most problems in high-dimension, we are allowed to fix a Gauge for one of these vectors (compatible with its distribution), since an average over all its possible realisations would just record the same contribution from each one of them. In this case, for simplicity, we choose $\boldsymbol{v} = \mathbb{I}$ to be a vector with all entries equal to 1 (this is still a vector on a sphere of radius N). Once this degree of freedom is fixed, the invariance is resolved so we still need to account for the average over the remaining vectors.

Let's define the partition function for the teacher vectors as:

$$Z_T = \int \boldsymbol{d\mu}(\boldsymbol{T}_+, \boldsymbol{T}_-) = \int \prod_{c=\pm} \left[d\mu\left(\boldsymbol{T}_c\right) \delta\left(|\boldsymbol{T}_c|^2 - N\tilde{Q} \right) \delta\left(\boldsymbol{T}_c \cdot \mathbb{I} - N\tilde{m}_c \right) \right] \delta\left(\boldsymbol{T}_+ \cdot \boldsymbol{T}_- - N\tilde{q} \right),$$

where the measures $\mu\boldsymbol{T}_\pm$ are in this case assumed to be factorised normal distributions. The Dirac's δ-functions ensure that the geometrical disposition of the model vectors is the one defined by the chosen magnetizations $\tilde{m}_\pm$ and the overlap $\tilde{q}$, and that the vectors are normalised to $\tilde{Q}_c$ (we are setting $\tilde{Q}_c = 1$).

At this point, and throughout this section, we use the integral representation of the δ-function:

$$\delta(x - aN) = \int \frac{d\hat{a}}{2\pi/N} e^{-i\hat{a}\left(\frac{x}{N} - a\right)}, \tag{D.6}$$

where $\hat{a}$ is a so-called conjugate field that plays a role similar to a Lagrange multiplier, enforcing the hard constraint contained in the δ-function. We can rewrite:

$$Z_T = \int \prod_{c=\pm} \frac{d\hat{\tilde{Q}}_c}{2\pi/N} \int \prod_{c=\pm} \frac{d\hat{\tilde{m}}_c}{2\pi/N} \int \frac{d\hat{\tilde{q}}}{2\pi/N} e^{N\Phi_T\left(\{\tilde{Q}_\pm, \tilde{m}_\pm, \tilde{q}\}, \{\hat{\tilde{Q}}_\pm, \hat{\tilde{m}}_\pm, \hat{\tilde{q}}\}\right)},$$

where the action Φ_T represents the entropy of configurations for the teacher that satisfy the chosen geometrical constraints. Given that the components of the teacher vectors are i.i.d., the entropy easily factorises over them. In high-dimensions, i.e. when $N \to \infty$, the integral will be dominated by "typical" configurations for the vectors, and the integral Z_T can be computed through a saddle-point approximation. We Wick rotate the fields in order to avoid dealing explicitly with imaginary quantities, and decompose $\Phi_T = g_{Ti} + g_{Ts}$:

$$g_{Ti} = -\left(\sum_c \hat{\tilde{m}}_c \tilde{m}_c + \sum_c \hat{\tilde{Q}}_c \tilde{Q}_c + \hat{\tilde{q}}\tilde{q} \right), \tag{D.7}$$

$$g_{Ts} = \log \int \mathcal{D}T_+ \int \mathcal{D}T_- \exp\left(\sum_c \hat{\tilde{Q}}_c T_c^2 + \sum_c \hat{\tilde{m}}_c T_c + \hat{\tilde{q}} T_+ T_- \right).$$

After a few Gaussian integrations the computation of the second term yields:

$$g_{Ts} = \frac{\left(1 - 2\hat{\tilde{Q}}_{-1}\right) \hat{\tilde{m}}_1^2 + \left(1 - 2\hat{\tilde{Q}}_1\right) \hat{\tilde{m}}_{-1}^2 + 2\hat{\tilde{q}}\hat{\tilde{m}}_1\hat{\tilde{m}}_{-1}}{2\left(\left(1 - 2\hat{\tilde{Q}}_{-1}\right)\left(1 - 2\hat{\tilde{Q}}_1\right) - \hat{\tilde{q}}^2\right)} - \frac{1}{2}\log\left(\left(1 - 2\hat{\tilde{Q}}_1\right)\left(1 - 2\hat{\tilde{Q}}_{-1}\right) - \hat{\tilde{q}}^2\right).$$

Now, in order to complete the computation of the partition function Z_T, we have impose the saddle point condition for Φ_T, which is realised when the entropy is extremised with respect to the fields we introduced. The saddle point equations for the teacher conjugate parameters give:

$$\partial_{\hat{\tilde{m}}_c} \Phi_T = 0 \to \tilde{m}_c = \frac{\left(1 - 2\hat{\tilde{Q}}_{-c}\right)\hat{\tilde{m}}_c + \hat{\tilde{q}}\hat{\tilde{m}}_{-c}}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)} \tag{D.8}$$

$$\partial_{\hat{\tilde{Q}}_c} \Phi_T = 0 \to \tilde{Q}_c = \left(1 - 2\hat{\tilde{Q}}_{-c}\right) \frac{\sum_{c'} \hat{\tilde{m}}_{c'}^2 \left(1 - 2\hat{\tilde{Q}}_{-c'}\right) + 2\hat{\tilde{q}}\hat{\tilde{m}}_+\hat{\tilde{m}}_-}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)^2} + \frac{\left(1 - 2\hat{\tilde{Q}}_{-c}\right) - \hat{\tilde{m}}_{-c}^2}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)} \tag{D.9}$$

$$\partial_{\hat{\tilde{q}}}\Phi_T = 0 \rightarrow \tilde{q} = \left(\frac{\hat{\tilde{q}} \sum_{c'} \hat{\tilde{m}}_c^2 \left(1 - 2\hat{\tilde{Q}}_{-c}\right) + 2\hat{\tilde{q}}^2 \hat{\tilde{m}}_+ \hat{\tilde{m}}_-}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)^2} \right) + \frac{\hat{\tilde{q}} + \hat{\tilde{m}}_+ \hat{\tilde{m}}_-}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)} \tag{D.10}$$

By moving around the terms in these equations one can find two identities that will be useful later in the computation:

$$\tilde{Q}_c - \tilde{m}_c^2 = \frac{\left(1 - 2\hat{\tilde{Q}}_{-c}\right)}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)} \tag{D.11}$$

$$\tilde{q} - \tilde{m}_+ \tilde{m}_- = \frac{\hat{\tilde{q}}}{\left(\left(1 - 2\hat{\tilde{Q}}_-\right)\left(1 - 2\hat{\tilde{Q}}_+\right) - \hat{\tilde{q}}^2\right)} \tag{D.12}$$

FREE ENTROPY OF THE LEARNING MODEL

In this section we aim to achieve analytical characterisation of typical learning performance in the T-M, i.e. to describe the solutions of the following optimisation problem:

$$\boldsymbol{w}_1^\star,\, \boldsymbol{w}_2^\star = \underset{\boldsymbol{w}_1,\, \boldsymbol{w}_2}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{w}_1,\, \boldsymbol{w}_2; \mathcal{D}), \tag{D.13}$$

where $\mathcal{D}$ represents a realisation of the data and $\mathcal{L}(\cdot)$ was defined in Eq. D.4. In typical statistical physics fashion, we can associate this problem with a Boltzmann-Gibbs probability measure, over the possible configurations of the student model parameters:

$$P(\boldsymbol{w}_1,\, \boldsymbol{w}_2; \mathcal{D}) = \frac{e^{-\beta \mathcal{L}(\boldsymbol{w}_1,\, \boldsymbol{w}_2; \mathcal{D})}}{Z_W}, \tag{D.14}$$

where the loss $\mathcal{L}$ plays the role of an the energy function, β is an inverse temperature and Z_W is the partition function (normalisation of the Boltzmann-Gibbs measure).

Since the loss is convex in the student parameters, when the inverse temperature is sent to infinity, $\beta \to \infty$, the probability measure focuses on the unique minimiser of the loss, representing the solution of the learning problem. In the asymptotic limit $N \to \infty$, the behaviour of this model becomes predictable since the overwhelming majority of the possible dataset realisations (with the same configuration of the generative parameters) will produce solutions with the same macroscopic properties (norm, test performance, etc). We therefore need to consider a self-averaging quantity, which is independent of the specific realisation of the dataset so that the typical learning scenario can be captured.

Thus, we compute the average free-energy:

$$\Phi_W = \lim_{N \to \infty} \lim_{\beta} \frac{1}{\beta N} \left\langle \log Z_W(\boldsymbol{w}_1, \boldsymbol{w}_2; \mathcal{D}_1, \mathcal{D}_2) \right\rangle_{\mathcal{D}_1, \mathcal{D}_2}. \tag{D.15}$$

This type of quenched average is not easily computed because of the $\log$ function in the definition. The replica trick, based on the simple identity $\lim_{n \to 0}(x^n - 1)/n = log(x)$, provides a method to tackle this computation. One can replicate the partition function, introducing n independent copies of the original system. Each of them, however, sees the same realisation of the data $\mathcal{D}$ (the "disorder" of the system, in the statistical physics terminology). When one takes the average over $\mathcal{D}$, the n replicas become effectively coupled, and can be intuitively interpreted as i.i.d. samples from the Boltzmann-Gibbs measure of the original problem. At the end of the computation, one takes the analytic continuation of the integer n to the real axis and computes the limit $\lim_{n \to 0}$, re-establishing the logarithm and the initial expression.

We start by working on the replicated volume (product over the n partition functions) $\Omega^n(\mathcal{D})$, which is still explicitly dependent on the sampled dataset:

$$\Omega^n(\mathcal{D}) = \int \frac{\boldsymbol{d\mu}(\boldsymbol{T}_+, \boldsymbol{T}_-)}{Z_T} \int \prod_{s,a} \left[db_s^a d\boldsymbol{w}_s^a e^{-\frac{\beta\gamma}{2} \|\boldsymbol{w}_1^a - \boldsymbol{w}_2^a\|^2} \prod_{\mu \in \mathcal{D}_s} e^{-\beta\, \ell\left(\frac{\boldsymbol{T}_{c^\mu} \cdot \boldsymbol{x}^\mu}{\sqrt{N}} + \tilde{b}_{c^\mu},\, \frac{\boldsymbol{w}_s^a \cdot \boldsymbol{x}^\mu}{\sqrt{N}} + b_s^a \right)} \right], \tag{D.16}$$

where $s = 1, 2$ indexes the two coupled student models and $a = 1, ..., n$ is the replica index.

To make progress we have to take the disorder average, i.e. the expectation over the distribution of $\boldsymbol{x}^\mu$ as defined in the T-M model. However, at this time the inputs appear inside the loss terms and taking a direct average is not feasible. We can exploit δ-functions in order to replace with dummy variables, u_μ and λ^a_μ, the dot products in the loss and isolate the input dependence in simpler exponential terms:

$$1 = \int \prod_\mu du_\mu \, \delta \left(u_\mu - \frac{\boldsymbol{T}_{c^\mu} \cdot \boldsymbol{x}^\mu}{\sqrt{N}} \right) \int \prod_{a,s,\mu \in \mathcal{D}_s} d\lambda^a_\mu \delta \left(\lambda^a_\mu - \frac{\boldsymbol{w}^a_s \cdot \boldsymbol{x}^\mu}{\sqrt{N}} \right) \tag{D.17}$$

$$= \int \prod_\mu \frac{du_\mu d\hat{u}_\mu}{2\pi} e^{i\hat{u}_\mu \left(u_\mu - \sum_{i=1}^N \frac{T_{c^\mu,i} x^\mu_i}{\sqrt{N}} \right)} \int \prod_{a,s,\mu \in \mathcal{D}_s} \frac{d\lambda^a_\mu d\hat{\lambda}^a_\mu}{2\pi} e^{i\hat{\lambda}^a_\mu \left(\lambda^a_\mu - \sum_{i=1}^N \frac{w^a_{s,i} x^\mu_i}{\sqrt{N}} \right)} \tag{D.18}$$

We can now evaluate the expectation over the input distribution, collecting all the terms where each given input appears. By neglecting terms that vanish in the $N \to \infty$ limit, for each pattern μ we get:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}^\mu} e^{-i \sum_a \hat{\lambda}^\mu_a \sum_{i=1}^N \frac{w^a_{s^\mu,i} x^\mu_i}{\sqrt{N}} - i\hat{u}_\mu \sum_{i=1}^N \frac{T_{c^\mu,i} x^\mu_i}{\sqrt{N}}} = \\ = &\prod_{i=1}^N e^{-ic^\mu \left(\sum_a \hat{\lambda}^\mu_a \frac{w^a_{s^\mu,i} v_i}{N} + \hat{u}^\mu \frac{T_{c^\mu,i} v^\mu_i}{N} \right)} \mathbb{E}_{z^\mu_i} e^{-i \left(\sum_a \hat{\lambda}^\mu_a \frac{w^a_{(\mu)i}}{\sqrt{N}} + \hat{u}_\mu \frac{T_{c^\mu,i}}{\sqrt{N}} \right) z^\mu_i} \end{aligned} \tag{D.19}$$

$$= e^{-ic^\mu \left(\sum_a \hat{\lambda}^\mu_a \frac{\sum_i w^a_{s^\mu,i}}{N} + \hat{u}^\mu \frac{\sum_i T_{c^\mu,i}}{N} \right) - \frac{\Delta_{c^\mu}}{2} \left(\sum_{ab} \hat{\lambda}^\mu_a \hat{\lambda}^\mu_b \frac{\sum_i w^a_{s^\mu,i} w^b_{s^\mu,i}}{N} + 2\hat{u}^\mu \sum_a \hat{\lambda}^\mu_a \frac{\sum_i w^a_{s^\mu,i} T_{c^\mu,i}}{N} + (\hat{u}^\mu)^2 \frac{\sum_i T^2_{c^\mu,i}}{N} \right)} . \tag{D.20}$$

To get Eq. D.20, we used the fact that the noise $\boldsymbol{z}^\mu$ is i.i.d. sampled from centred Gaussians of variance determined by the group, and explicitly used our Gauge choice $\boldsymbol{v} = \mathbb{I}$. In this expression we see appearing the relevant order parameters of the model, describing the overlaps between the student vectors, the shift vector and the teacher vectors. We are thus going to introduce via δ-functions the following parameters:

- $m^a_s = \frac{\boldsymbol{w}^a_s \cdot \mathbb{I}}{N}$, $\tilde{m}_c = \frac{\boldsymbol{T}_c \cdot \mathbb{I}}{N}$: magentisations in the direction of the $+$ group centre of the students and the teachers.
- $q^{ab}_s = \frac{\sum_i w^a_{si} w^b_{si}}{N}$: self-overlap between different replicas of each student.
- $R^a_{sc} = \frac{\sum_i w^a_{s,i} T_{c,i}}{N}$: overlap between student and teacher vectors.
- $\tilde{Q}_c = \frac{\sum_i T^2_{ci}}{N}$: norm of the teacher vectors ($= 1$ by assumption).

After the introduction of these order parameters (via the integral representation of the δ-function) the replicated volume can be expressed as:

$$\Omega^n = \int \prod_{s,a} \frac{dm^a_s d\hat{m}^a_s}{2\pi/N} \int \prod_{sc,a} \frac{dR^a_{sc} d\hat{R}^a_{sc}}{2\pi/N} \int \prod_{s,ab} \frac{dq^{ab}_s d\hat{q}^{ab}_s}{2\pi/N} \int \prod_c db^a_c G^N_I G^N_S \prod_{sc} G_E(s,c)^{\alpha_{c,s} N} \tag{D.21}$$

where $\alpha_{c,s} N$ indicates the number of patterns from group c contained in the data slice $\mathcal{D}_s$ given to student s. We also introduced the interaction, the entropic and the energetic terms:

$$G_I = \exp \left(-N \left(\sum_{s,a} \hat{m}^a_s m^a_s + \sum_{s,ab} \hat{q}^{ab}_s q^{ab}_s + \sum_{sc,a} \hat{R}^a_{sc} R^a_{sc} \right) \right) \tag{D.22}$$

$$\begin{aligned} G_S = &\int \prod_c \mathcal{D}T_c \exp \left(\sum_c \hat{\tilde{Q}}_c T^2_c + \sum_c \hat{\tilde{m}}_c T_c + \hat{\tilde{q}} T_+ T_- \right) \\ \times &\int \prod_{s,a} d\mu \left(w^a_s \right) e^{-\beta\gamma (w^a_1 - w^a_2)^2} \exp \left(\sum_{s,a} \hat{m}^a_s w^a_s + \sum_{s,ab} \hat{q}^{ab}_s w^a_s w^b_s + \sum_{sc,a} \hat{R}^a_{sc} w^a_s T_c \right) \end{aligned} \tag{D.23}$$

$$G_E(s,c) = \int \frac{du d\hat{u}}{2\pi} e^{iu\hat{u}} \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a} \right) e^{-\frac{\Delta^c}{2} \sum_{ab} \hat{\lambda}_a \hat{\lambda}_b q_s^{ab} - \Delta^c \hat{u} \sum_a \hat{\lambda}_a R_{sc}^a - \frac{\Delta^c}{2} (\hat{u})^2 \tilde{Q}_c}$$
$$\times \prod_a e^{-\beta\, \ell\left(u + c\tilde{m}_c + \tilde{b}_c, \lambda^a + c m_s^a + b_s^a\right)} \tag{D.24}$$

The shorthand notation $\mathcal{D}x = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ is used to indicate a normal Gaussian measure.

REPLICA SYMMETRIC ANSATZ

To make further progress, we have to make an assumption for the structure of the introduced order parameters. Given the convex nature of the optimisation objective D.4, the simplest possible ansatz, the so-called replica symmetric (RS) ansatz, is fortunately exact. Replica symmetry introduces a strong constraint for the overlap parameters, requiring the n replicas of the students to be indistinguishable and the free entropy to be invariant under their permutation. Mathematically, the RS ansatz implies that:

- $m_s^a = m_s$ for all $a = 1, ..., n$ (same for the conjugate)
- $R_{sc}^a = R_{sc}$ for all $a = 1, ..., n$ (same for the conjugate)
- $q_s^{ab} = q_s$ for all $a > b$, $q_s^{ab} = Q_s$ for all $a = b$ (same for the conjugate)
- $b_s^a = b_s$ for all $a = 1, ..., n$

Moreover, since we want to describe the minimisers of the loss, we are going to take the $\beta \to \infty$ limit in the Gibbs-Boltzmann measure. The replicas, which represent independent samples from it, will collapse on the unique minimum. This is represented by the following scaling law with β for the order parameters, which will be used below:

$$Q - q = \delta q/\beta; \quad \hat{Q} - \hat{q} = -\beta\delta\hat{q}; \quad \hat{q} \sim \beta^2 \hat{q}; \quad \hat{m} \sim \beta\hat{m}; \quad \hat{R} \sim \beta\hat{R} \tag{D.25}$$

INTERACTION TERM

We now proceed with the calculation of the different terms in D.21, where we can substitute the RS ansatz. In the interaction term, neglecting terms of $\mathcal{O}(n^2)$, we get:

$$G_i = \exp\left(-n \left(\sum_s \left(\hat{m}_s m_s + \sum_c \hat{R}_{sc} R_{sc} + \frac{\hat{Q}_s Q_s}{2} - \frac{\hat{q}_s q_s}{2} \right) \right) \right) \tag{D.26}$$

In the $\beta \to \infty$ limit the expression becomes:

$$\log(G_i)/n = g_i = -\beta \left(\sum_s \left(\hat{m}_s m_s + \sum_c \hat{R}_{sc} R_{sc} + \frac{1}{2} \left(\hat{q}_s \delta q_s - \delta\hat{q}_s q_s \right) \right) \right) \tag{D.27}$$

ENTROPIC TERM

In the entropic term the computation is more involved, due to the couplings between the Gaussian measures for the teachers and for those of the students. We substitute the RS ansatz in expression D.23 to get:

$$G_S = \int \mathcal{D}T_+ \int \mathcal{D}T_- \exp\left(\sum_c \hat{\tilde{Q}}_c T_c^2 + \sum_c \hat{\tilde{m}}_c T_c + \hat{\tilde{q}} T_+ T_- \right) \int \prod_{s,a} d\mu\left(w_s^a\right) e^{-\frac{\gamma}{2}(w_1^a - w_2^a)^2}$$
$$\times \prod_s \exp\left(\hat{m}_s \sum_a w_s^a + \frac{1}{2}\left(\hat{Q}_s - \hat{q}_s \right) \sum_a \left(w_s^a\right)^2 + \frac{1}{2}\hat{q}_s \left(\sum_a w_s^a \right)^2 + \sum_c \hat{R}_{sc} \sum_a w_s^a T_c \right) \tag{D.28}$$

We perform a Hubbard-Stratonovich transformation to remove the squared sum in the previous equation, introducing the Gaussian fields z_s. Then, we rewrite coupling term between the teachers as $\hat{\tilde{q}}T_+T_- = \frac{\hat{\tilde{q}}}{2}(T_+ + T_-)^2 - \frac{\hat{\tilde{q}}}{2}(T_+^2 + T_-^2)$, and perform a second Hubbard-Stratonovich transformation, with field $\tilde{z}$, to remove the explicit coupling between T_+ and T_-. Similarly, the elastic coupling between the students can be turned into a linear term with fields z_{12}^a:

$$= \int \mathcal{D}\tilde{z} \int \prod_s \mathcal{D}z_s \int \frac{dT_c}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_c \left(1 - 2\hat{\tilde{Q}}_c + \hat{\tilde{q}}\right) T_c^2 + \sum_c \left(\hat{\tilde{m}}_c + \sqrt{\hat{\tilde{q}}}\tilde{z}\right) T_c\right) \int \prod_a \mathcal{D}z_{12}^a$$
$$\times \int \prod_{s,a} d\mu\left(w_s^a\right) \prod_s \exp\left(\frac{1}{2}\left(\hat{Q}_s - \hat{q}_s\right)\sum_a \left(w_s^a\right)^2 + \left(\hat{m}_s + \sum_c \hat{R}_{sc}T_c + \sqrt{\hat{q}_s}z_s + is\sqrt{\gamma}z_{12}^a\right)\sum_a w_s^a\right) \tag{D.29}$$

After rescaling the variances of the teacher measures and centring them, one can factorise over the replica index and take the $n \to 0$ limit, obtaining the following expression for $g_S = \log G_S/n$:

$$g_S = A + \int \prod_s \mathcal{D}z_s \int \prod_c \mathcal{D}T_c \int \mathcal{D}\tilde{z} \log \int \mathcal{D}z_{12} \int \prod_s d\mu\left(w_s\right) \exp\left(\frac{1}{2}\left(\hat{Q}_s - \hat{q}_s\right) w_s^2 + B_s w_s\right) \tag{D.30}$$

where:

$$A = \frac{\sum_c \hat{\tilde{m}}_c^2 \left(1 - 2\hat{\tilde{Q}}_{-c}\right) + 2\hat{\tilde{q}}\left(\sum_c \hat{\tilde{m}}_c\right)^2}{2\left(\left(1 - 2\hat{\tilde{Q}}_+\right)\left(1 - 2\hat{\tilde{Q}}_-\right) - \hat{\tilde{q}}^2\right)} \tag{D.31}$$

$$B_s = b_s\left(T_\pm, z_\pm, \tilde{z}, z_s\right) + is\sqrt{\gamma}z_{12} \tag{D.32}$$

$$b_s = \hat{m}_s + \sqrt{\hat{q}_s}z_s + \sum_c \left[\tilde{m}_c\hat{R}_{sc} + \frac{\hat{R}_{sc}}{\sqrt{\left(1 - 2\hat{\tilde{Q}}_c + \hat{\tilde{q}}\right)}}T_c + \frac{\sqrt{\hat{\tilde{q}}}}{\sqrt{1 - \sum_{c'} \frac{\hat{\tilde{q}}}{\left(1 - 2\hat{\tilde{Q}}_{c'} + \hat{\tilde{q}}\right)}}} \frac{\hat{R}_{sc}}{\left(1 - 2\hat{\tilde{Q}}_c + \hat{\tilde{q}}\right)}\tilde{z}\right] \tag{D.33}$$

In the $\beta \to \infty$ limit, and considering the L_2-regularisation on the student weights $d\mu\left(w\right) = \frac{dw}{\sqrt{2\pi}}e^{-\frac{\beta\lambda}{2}w^2}$ we get:

$$g_S = A + \int \prod_s \mathcal{D}z_c \int \prod_c \mathcal{D}T_c \int \mathcal{D}\tilde{z} \log \int \mathcal{D}z_{12} \exp\left(\sum_s \max_{w_s}\left(-\frac{\lambda + \delta\hat{q}_s}{2}w_s^2 + B_s w_s\right)\right) \tag{D.34}$$

and the maximisation gives:

$$w_s^\star = \frac{B_s}{\left(\lambda + \delta\hat{q}_s\right)}; \quad \max_{w_s}\left(-\frac{\lambda + \delta\hat{q}_s}{2}w_s^2 + B_s w_s\right) = \frac{B_s^2}{2\left(\lambda + \delta\hat{q}_s\right)} \tag{D.35}$$

Substituting the above described scaling laws for the order parameters in the $\beta \to \infty$ limit one finds that the A term becomes sub-dominant and can be ignored. The remaining steps are quite tedious, but the procedure to obtain the final result for the entropic channel is straightforward:

- Expand the sums in Eq.D.34.
- Perform the z_12 Gaussian integration and take the $\log$ of the result.
- Identify the terms that have even powers in the Hubbard-Stratonovich Gaussian fields and in the teacher variables. The Gaussian integrations will kill all the remaining cross terms, so they can be ignored.
- Perform the remaining Gaussian integrations.
- Use identities D.8, D.11 and D.12 to remove the dependence on the conjugate fields appearing in the Teacher measure and only retain a dependence on $\tilde{m}_c$, $\tilde{Q}_c$, and $\tilde{q}$.

The final expression reads:

$$g_S = \frac{\beta}{2\left(\prod_s(\lambda + \gamma + \delta\hat{q}_s) - \gamma^2\right)} \left[\left(\sum_s \left(\hat{m}_s + \sum_s \tilde{m}_c\hat{R}_{sc}\right)^2 (\lambda + \gamma + \delta\hat{q}_{\neg s}) + 2\gamma \prod_s \left(\hat{m}_s + \sum_c \tilde{m}_c\hat{R}_{sc}\right)\right)\right. \tag{D.36}$$

$$+\left(\sum_{s}\hat{q}_s\left(\lambda+\gamma+\delta\hat{q}_{\neg s}\right)\right)+\left(\sum_{c}\left(\tilde{Q}_c-\tilde{m}_c^2\right)\left(\sum_{s}\hat{R}_{sc}^2\left(\lambda+\gamma+\delta\hat{q}_{\neg s}\right)+2\gamma\prod_{s}\hat{R}_{sc}\right)\right) \tag{D.37}$$

$$+\left(2\left(\tilde{q}-\tilde{m}_+\tilde{m}_-\right)\left(\sum_{s}\left(\prod_{c}\hat{R}_{sc}\left(\lambda+\delta\hat{q}_{\neg s}\right)\right)+\gamma\left(\prod_{c}\left(\sum_{s}\hat{R}_{sc}\right)\right)\right)\right)\Bigg] \tag{D.38}$$

ENERGETIC TERM

We can compute the energetic channel for a generic student s and a generic data group c. Each term will be multiplied by $\alpha_{c,s}$, determining the fraction of inputs from group c in the dataset $\mathcal{D}_s$ of student s. For simplifying the notation in this section we drop the indices s, c, with the understanding that the all the order parameters, and model parameters, appearing in the following expressions are those corresponding to a specific pair of these indices.

Substituting the RS ansatz in Eq. D.24 we get:

$$G_E=\int\frac{dud\hat{u}}{2\pi}e^{iu\hat{u}}\int\prod_{a}\left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi}e^{i\lambda^a\hat{\lambda}^a}\right)e^{-\frac{\Delta}{2}\sum_{ab}\hat{\lambda}_a\hat{\lambda}_b q-\Delta\hat{u}R\sum_a\hat{\lambda}_a-\frac{\Delta}{2}(\hat{u})^2\tilde{Q}} \tag{D.39}$$

$$\times\prod_{a}e^{-\beta\,\ell\left(u+c\tilde{m}+\tilde{b},\lambda^a+cm+b\right)} \tag{D.40}$$

We can start by evaluating the Gaussian in $\hat{u}$, then performing a Hubbard-Stratonovich transformation, with field z, to remove the squared sums on the replica index. Following up with the Gaussian integration in $\hat{\lambda}$ we find that the argument of the integrations factorises over the replica index. Up to first order in n when $n\to 0$, we find for $g_E=\log G_E/n$:

$$g_E=\int\mathcal{D}z\int\mathcal{D}u\log\int\mathcal{D}\lambda e^{-\beta\,\ell\left(\sqrt{\Delta\tilde{Q}}u+c\tilde{m}+\tilde{b},\sqrt{\Delta(Q-q)}\lambda+\frac{\sqrt{\Delta}R}{\sqrt{\tilde{Q}}}u+\sqrt{\Delta\frac{(q-R^2)}{\tilde{Q}}}z+cm+b\right)} \tag{D.41}$$

and in the the $\beta\to\infty$ limit we can solve the integral by saddle-point:

$$\log\int\mathcal{D}\lambda e^{-\beta\,\ell\left(\sqrt{\Delta\tilde{Q}}u+c\tilde{m}+\tilde{b},\sqrt{\Delta(Q-q)}\lambda+\frac{\sqrt{\Delta}R}{\sqrt{\tilde{Q}}}u+\sqrt{\Delta\frac{(q-R^2)}{\tilde{Q}}}z+cm+b\right)}=-\beta M \tag{D.42}$$

with:

$$M=\min_{\lambda}\frac{\lambda^2}{2}+\ell\left(\sqrt{\Delta\tilde{Q}}u+c\tilde{m}+\tilde{b},\sqrt{\Delta\delta q}\lambda+\frac{\sqrt{\Delta}R}{\sqrt{\tilde{Q}}}u+\sqrt{\Delta\frac{(q-R^2)}{\tilde{Q}}}z+cm+b\right) \tag{D.43}$$

To simplify further, we can shift $\frac{\sqrt{\Delta}R}{\sqrt{\tilde{Q}}}u+\sqrt{\Delta\frac{(q-R^2)}{\tilde{Q}}}z\to\sqrt{\Delta q}z'$. Then, given the definition of the logistic loss D.5, we can split the u integration over the intervals $\sqrt{\Delta\tilde{Q}}u+c\tilde{m}_c>0$ and $\sqrt{\Delta\tilde{Q}}u+c\tilde{m}_c<0$ and eventually get (re-establishing the s, c indices):

$$g_E(s,c)=\sum_{y}\int\mathcal{D}zH\left(-y\frac{q_s\frac{c\tilde{m}_c+\tilde{b}_c}{\sqrt{\tilde{Q}_c}}+\sqrt{\Delta_c}R_{sc}z}{\sqrt{\Delta_c\left(q_s-R_{sc}^2\right)}}\right)M_E\left(y,s,c\right) \tag{D.44}$$

Where $H(x)=\frac{1}{2}\operatorname{erfc}(x/\sqrt{2})$ is the Gaussian tail function and we defined the proximal:

$$M_E\left(y,s,c\right)=\max_{\lambda}-\frac{\lambda^2}{2}-\ell\left(y,\sqrt{\Delta_c\delta q_s}\lambda+\sqrt{\Delta_c q_s}z+cm_s+b_s\right) \tag{D.45}$$

Note that this simple 1D optimisation problem has to be solved numerically in correspondence of each point evaluated in the integral.

The reweighing strategy is easily embedded in this calculation by explicitly changing the definition of ℓ, adding a different weight $W_{c,y}$ for each combination of label and group membership. Defining a one-hot encoding vector for the teacher-produced label, $Y \in \mathbb{R}^2$, and a output probability (constructed from the sigmoid function) for the student, $P(\hat{Y})$, the reweighed cross-entropy loss can be written as:

$$\mathcal{L}(\mathcal{D}) = \sum_{c=\pm} \sum_{y=0,1} (W)_{(c,y)} Y_y \log P(\hat{Y}_y). \tag{D.46}$$

For the sake of simplicity we reduced the degrees of freedom to two, parameterising these weights as:

$$W = 2 \begin{pmatrix} w_+ w_1 & w_+(1-w_1) \\ (1-w_+)w_1 & (1-w_+)(1-w_1) \end{pmatrix} \tag{D.47}$$

where $w_+, w_1 \in [0,1]$ can be used to increase the relative weight of a misclassification errors in the group $+$ and label 1 respectively.

Different losses could be chosen instead of the cross-entropy and, again, only the numerical optimisation of the proximal would be affected.

SADDLE-POINT OF THE FREE-ENTROPY

We thus have found that the free-entropy Φ_W can be written as a simple function of few scalar order parameters. In the high-dimensional limit, the integral in D.21 is dominated by the typical configuration of the order parameters, which is found by extremising the free-entropy with respect to all the overlap parameters:

$$\Phi_W = \underset{o.p.}{\text{extr}} \left\{ g_I + g_S + \sum_{s,c} \alpha_{s,c}\, g_E(s,c) \right\} \tag{D.48}$$

The saddle-point is typically found by fixed-poimnt iteration: setting each derivative, with respect to the order parameters, to zero returns a saddle-point condition for the conjugate parameters, and vice-versa.

The fixed-point is uniquely determined by the value of the generative parameters $\tilde{m}_\pm$,$\tilde{Q}_\pm$,$\tilde{q}$ and the pattern densities $\alpha_{s,c}$. In the main text, for simplicity we parameterise $\alpha_{s,c}$ through the fraction η, which represents the percentage of patterns from group $+$ assigned to the first student model.

The special case of a single student model is obtained from this calculation by setting $\gamma = 0$ and assigning all the inputs in the first dataset $\mathcal{D}_1$.

TEST ACCURACY

All the performance assessment metrics employed in this paper can be derived from the confusion matrix, which measures the TP, FP, TN, FN rates on new samples from the T-M. These quantities can be evaluated analytically and are easily expressed as a function of the saddle-point order parameters obtained in the previous paragraphs.

Suppose we obtain a new data point with label y from group c. The probability of obtaining an output $\hat{y}$ from the trained model s is given by:

$$P\left(Y = y, \hat{Y} = \hat{y}\right) = \mathbb{E}_{\mathbf{x}(c)} \left\langle \Theta\left(y \left(\frac{\mathbf{T}_c \cdot \mathbf{x}(c)}{\sqrt{N}} + \tilde{b} \right)\right) \Theta\left(\hat{y} \left(\frac{\mathbf{w}_s \cdot \mathbf{x}(c)}{\sqrt{N}} + b \right)\right) \right\rangle_{\mu(\boldsymbol{T},\boldsymbol{w})} \tag{D.49}$$

$$= \mathbb{E}_{\mathbf{x}(c)} \left\langle \int \frac{du\,d\hat{u}}{2\pi} e^{i\hat{u}\left(u - \sum_{i=1}^N \frac{T_i x_i}{\sqrt{N}}\right)} \int \frac{d\lambda\, d\hat{\lambda}}{2\pi} e^{i\hat{\lambda}\left(\lambda - \sum_{i=1}^N \frac{w_i x_i}{\sqrt{N}}\right)} \right\rangle \Theta\left(y \left(u + \tilde{b} \right)\right) \Theta\left(\hat{y} \left(\lambda + b \right)\right) \tag{D.50}$$

where, following the same lines as in the free-entropy computation, we used δ-functions to extract the dependence on the input, to facilitate the expectation:

$$\mathbb{E}_{\boldsymbol{x}(c)} \left\langle e^{-i\hat{\lambda} \frac{\boldsymbol{w}_s \cdot \boldsymbol{x}(c)}{\sqrt{N}} - i\hat{u} \frac{\mathbf{T} \cdot \mathbf{x}(c)}{\sqrt{N}}} \right\rangle \tag{D.51}$$

$$= e^{-ic\left(\hat{\lambda} m + \hat{u}\tilde{m}\right)} e^{-\frac{\Delta}{2}\left(\hat{\lambda}^2 Q + 2\hat{u}\hat{\lambda} R + \hat{u}^2 \tilde{Q}\right)}. \tag{D.52}$$

We have substituted the overlaps that come out of the average with their typical values in the Boltzmann-Gibbs measure of the T-M. Note that we can substitute $q = Q$ since they are equal to first order in the $\beta \to \infty$ limit.

The Gaussian integrals can be computed and one gets the final expression:

$$P\left(Y = y, \hat{Y} = \hat{y}\right) = \int_{-\infty}^{\infty} \mathcal{D}u \Theta\left(y\left(\sqrt{\Delta_c}u + c\tilde{m}_c + \tilde{b}_c\right)\right) H\left(-\hat{y}\frac{\sqrt{\Delta_c}R_{sc}u + cm_s + b_s}{\sqrt{\Delta_c\left(q_s - R_{sc}^2\right)}}\right) \tag{D.53}$$

Similarly, one can also obtain e.g. the label 1 frequency:

$$P\left(Y = 1\right) = \rho H\left(-\frac{\tilde{m}_+ + \tilde{b}_+}{\sqrt{\Delta_+\tilde{Q}_+}}\right) + \left(1 - \rho\right) H\left(\frac{\tilde{m}_- - \tilde{b}_-}{\sqrt{\Delta_-\tilde{Q}_-}}\right) \tag{D.54}$$

and the generalisation error:

$$\epsilon_g = \int_{-\infty}^{\infty} \mathcal{D}u H\left(\text{sign}\left(\left(\sqrt{\Delta_c}u + c\tilde{m}_c + \tilde{b}_c\right)\right)\frac{\sqrt{\Delta_c}R_{sc}u + cm_s + b_s}{\sqrt{\Delta_c\left(q_s - R_{sc}^2\right)}}\right). \tag{D.55}$$

E EXPLORATION OF THE PARAMETER SPACE

E.1 PARAMETERS USED IN THE FIGURES

This section present a list of the parameters used in for the T-M model in the figures of the paper. Some of the parameters are already discussed in the figures' captions, so this list will characterise only the remaining parameters:

- Fig. 1 (*center*): $\Delta_+ = 0.5, \Delta_- = 2_0.5, \alpha = 2.5, q_T - 0.2$.
- Fig. 2 (first two panels): $m_\pm = 0.2$, $\alpha = 0.5, \Delta_+ = 0.5, \Delta_- = 0.5, b_+ = 0, b_- = 0$.
- Fig. 2 (last two panels): $\alpha = 0.5$, $\Delta_+ = 0.5, \Delta_- = 0.5, b_+ = 0, b_- = 0$.
- Fig. 3: $\alpha = 0.5, q_T = 1, m = 0.5, b_+ = 0, b_- = 0$.
- Fig. 4: $\rho = 0.1, m = 0.2, \Delta_+ = 0.5, \Delta_- = 0.5, b_+ = 0, b_- = 0$.
- Fig. 5 and Fig. 6 (first two panels): $\rho = 0.1, q_T = 0.8, \Delta_+ = 2.0, \Delta_- = 0.5, \alpha = 0.5, m_+ = 0.3, m_- = 0.1, b_+ = 0.5, b_- = 0.5$.

E.2 SUPPORTING RESULTS

This section presents supporting results on the sources of bias. In Fig. E.1, we re-propose the the study of the disparate impact (DI) depending on the relative representation ρ and the rule similarity q_T, paying close attention to the role of the group-label correlation m_+, m_-. Interestingly, if $m_+ = m_- = 0$, when the rules become identical ($q_T = 1$) the bias is removed. However if $m_+ = m_- \neq 0$ this is no longer true. This shows once again that it is not sufficient for a classifier to be able of reproducing the rule, as bias can appear in reason of other concurring factors.

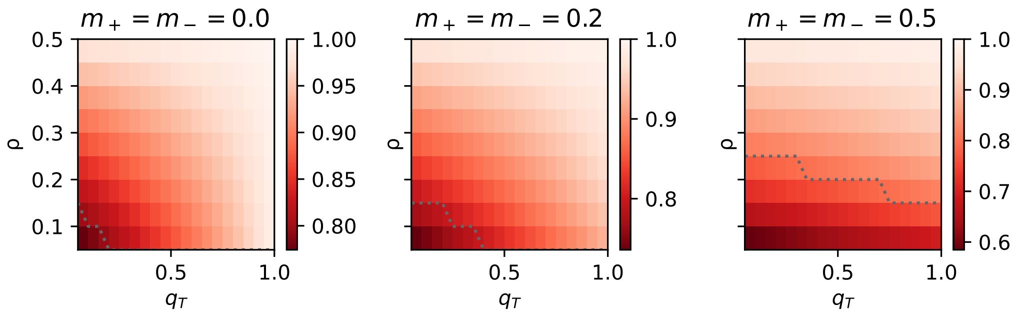

Figure E.1: **Bias with two different rules to be learned.** The three phase diagrams give the DI depending on ρ (y-axis) and q_T (x-axis). Moving from the left panel to the right panel m_+ and m_i are increased. The other parameters are: $\alpha = 0.5, \Delta_+ = 0.5, \Delta_- = 0.5, b_+ = 0, b_- = 0$.

The main difference with respect to the case with $q_T \neq 1$ is that, if $q_T = 1$, increasing the amount of training data can be a solution. In fact, bias at $q_T = 1$ is due to overfitting with respect ot the largest sub-population, and this effect can be cured by increasing in α. This is illustrated in Fig. E.2, that extends the figure of the main text showing the effect of α. Moving from left to right, α increases and the area where the 80% rule is violated shrinks down.

The results shown until this point are agnostic with respect to the relative fraction of labels inside the sub-populations. When this quantity is strongly varied across the groups, it can contribute to an additional source of bias, especially if combined with a small relative representation. Indeed, the classifier can simply bias its prediction towards the most likely outcome reaching an accuracy that apparently exceeds random guessing, without effectively doing any informed prediction. Many factors play a role in deciding the relative fraction of labels in the T-M model, the bias terms (b_+ and b_-) are the most relevant since they directly shift the decision boundaries. We consider these two parameters in Fig. E.3 to exemplify this concept.

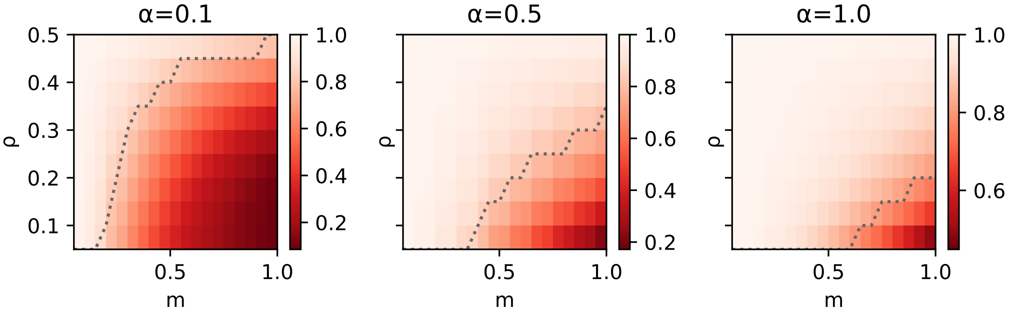

Figure E.2: **Bias with a learnable rule.** We show the accuracy gain as function of the proportion of group + (ρ) and the correlation between label and group (m_+, m_-). The different figures show how of increasing the dataset size (increasing from left to right) mitigates the bias. The other parameters are: $q_T = 1.0, \Delta_+ = 0.5, \Delta_- = 0.5, b_+ = 0, b_- = 0$.

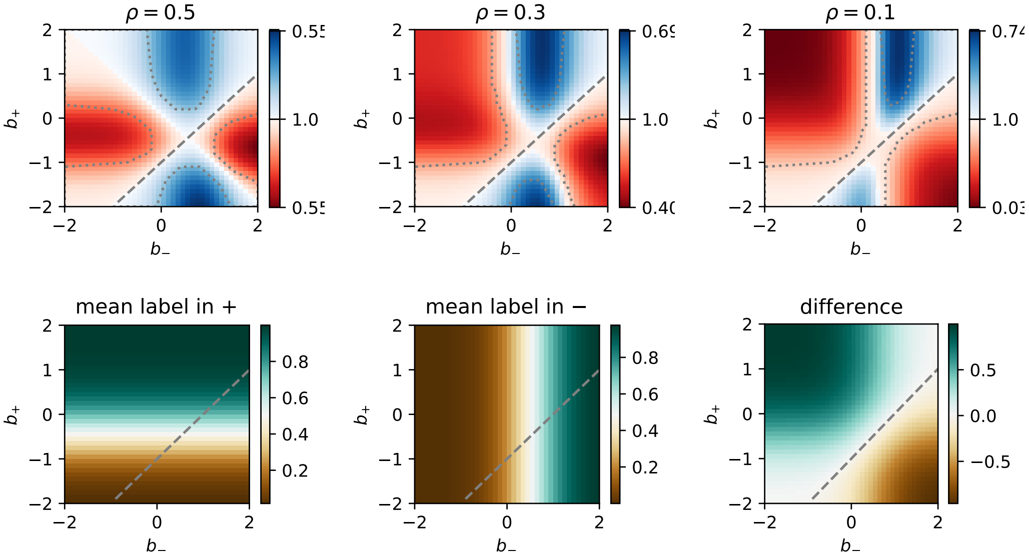

Figure E.3: **Labels within groups and classifier bias.** The *first row* shows the DI as faction of b_+ and b_- with $\Delta_+ = \Delta_- = 0.5$, $\alpha = 0.5$, $m_+ = m_- = 0.5$. From left to right, the relative representation ρ moves from equally represented groups to having group + under-represented. The 80% threshold is denoted by the dotted line. The dashed line indicates equal within-group label fraction. The *second row* shows the average labelling in + (left), − (centre), and their difference (right). Notice that these diagrams are independent of ρ and therefor apply to the three settings shown in the first row.

When the sub-populations are equally represented $\rho = 0.5$, the separations between bias towards + or − is clearly marked by two straight lines. One separation is simply given by the line of equal label fraction, the other is given by the uncertainty of the classifier, receiving contrasting inputs from the two groups. As the relative representation ρ decreases, the classifier accommodates the inputs from the largest group and the separation line is distorted. Finally, observe that the line of equal label fraction (bottom right panel) is not centred in the diagram because $m_+ = m_- \neq 0$.

F MITIGATION STRATEGIES

In this section we summarise few well-established criteria in Table 1. The common denominator among them is that they make a requirement of independence between the distribution of a certain type of prediction outcome and the group-membership. From a probabilistic point of view, denoting with E an random variable that represents the outcome, and with C the group membership, the classifier has to satisfy $P(E|C) = P(E)$. For instance, *Equal Opportunity* requires the ML model to achieve equal true positive rates, independent of the sub-population.

FAIRNESS METRIC	**CONDITION**
Statistical Parity	$\mathbb{P}[\hat{Y} = y \vert C = c] = \mathbb{P}[\hat{Y} = y] \ \forall y, c$
Equal Opportunity	$\mathbb{P}[\hat{Y} = 1 \vert C = c, Y = 1] = \mathbb{P}[\hat{Y} = 1 \vert Y = 1] \ \forall c$
Equal Accuracy	$\mathbb{P}[\hat{Y} = y \vert C = c, Y = y] = \mathbb{P}[\hat{Y} = y \vert Y = y] \ \forall y, c$
Equal Odds	$\mathbb{P}[\hat{Y} = 1 \vert C = c, Y = 1] = \mathbb{P}[\hat{Y} = 1 \vert Y = 1] \cap$ $\mathbb{P}[\hat{Y} = 1 \vert C = c, Y = 0] = \mathbb{P}[\hat{Y} = 1 \vert Y = 0] \ \forall c$
Predicted Parity	$\mathbb{P}[Y = 1 \vert C = +, \hat{Y} = y] = \mathbb{P}[Y = 1 \vert C = -, \hat{Y} = y]$ $= \mathbb{P}[Y = 1 \vert \hat{Y} = y] \ \forall y$

Table 1: **List of Fairness Metrics.** *Statistical Parity*: Equal fractions of each group should be treated as belonging to the positive class Dwork et al. (2012); Kleinberg et al. (2016); Corbett-Davies et al. (2017). *Equal Opportunity*: Each group need to achieve equal true positive rateHardt et al. (2016). *Equal Accuracy*: Each group is required to achieve the same level of accuracy. *Equal Odds*: Each group should achieve equal true positive and false positive ratesFeldman et al. (2015); Zafar et al. (2017b). *Predicted Parity*. Given the inputs are classified with label y, the fraction of input with true label y^* should be consistent across sub-populations. This gives two methods: *predicted parity 1* requires the condition only for $y^* = 1$, while predicted parity 10 requires the condition for both $y^* = 1$ and $y^* = 0$ Chouldechova (2017).

F.1 ADDITIONAL RESULTS ON MITIGATION STRATEGIES

Some strategies require information concerning the group membership of each data point. Depending on the situation, this information may contain errors or it may even be unavailable. Consequently we should take into account the robustness of the mitigation strategies with respect to these errors. Call η the fraction of points for which the group was correctly assessed. The phase diagrams in Fig. F.3a show the DI under the reweighing mitigation scheme (controlling the group importance in the loss) and the coupled classifier mitigation. We can clearly observe a greater resilience to the error rates in the case of our strategy. The reweighing strategy appears to have low DI only in extreme cases, where the accuracy on the largest sub-population is greatly deteriorated.

We can understand the larger picture by looking at the different fairness metrics described in the main text, Fig. F.3b, for which the same observations apply. Since η is not an actual hyper-parameter, but rather represents an imperfect imputation of the group structure, we consider the maximum for each value of η. The picture seems quite robust on the side of reweighing (upper group): for every η the maximum is achieved for different values of the parameters. Instead, the picture changes for the coupled classifiers (lower group): the method is robust to this perturbation until a critical value (roughly 25% of mismatched inputs), where the minima of the MI become inconsistent and therefore the fairness metrics cannot be optimised all at once.

Validation of re-weighting result. In the main text we show the effect of reweighing in the synthetic model. The same analysis can be applied to real data, yielding similar results. In particular, in line with the other validations, we present in Fig. F.4 the result for the CelebA dataset when the splitting is done according to the ”Wearing_Lipstick” and the target feature is ”Wavy_Hair”.

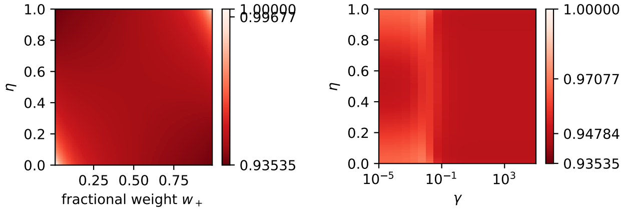

Figure F.1: **DI with errors in the group membership.**

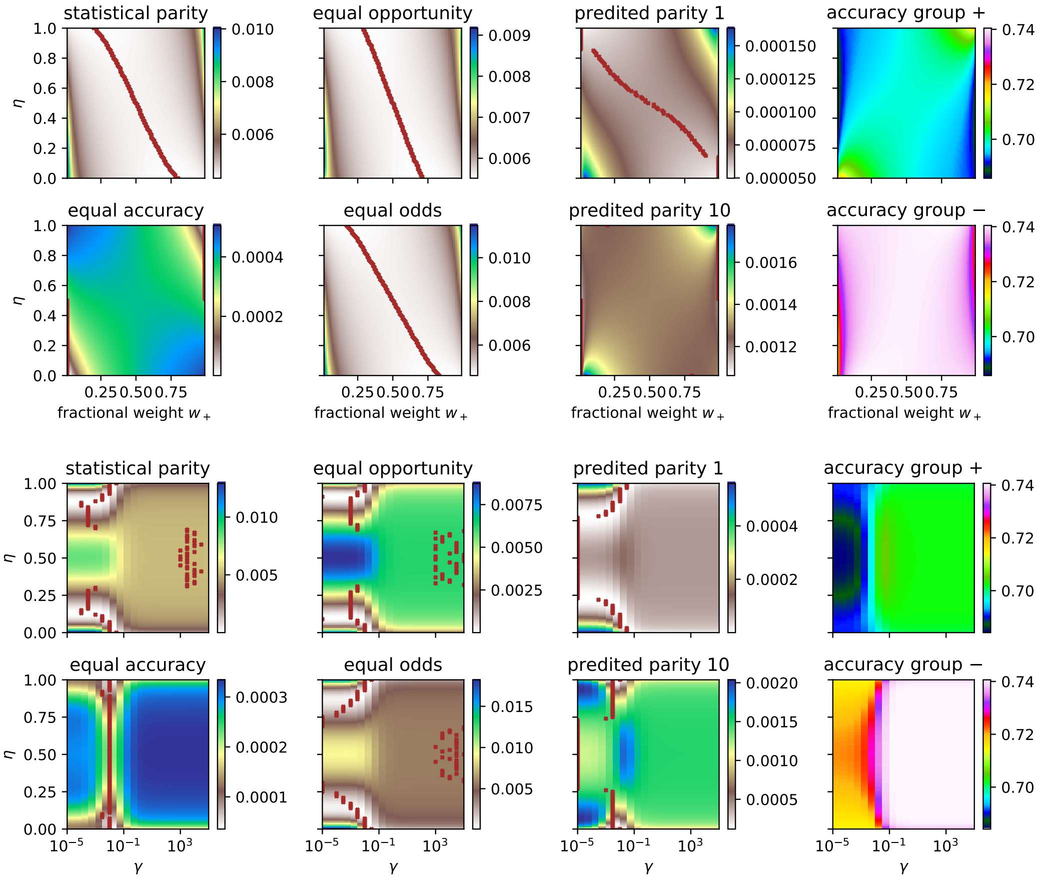

Figure F.2: **MI with errors in the group membership.**

Figure F.3: **(a)** The diagrams show the DI for the re-weighting according to membership (left) and the coupled classifiers (right). The colormaps are matched: the maximum is set to 1 (indicating absence of bias), the minimum is given by minimum DI registered by the two methods (i.e. the one of re-weighting). We also add ticks on the two colormaps, to indicate the extremes achieved by the two methods. **(b)** The upper group refers to re-weighting in the subpopulations, the lower group refers to the coupled classifiers. Refer to Fig.6 of the main text for more details.

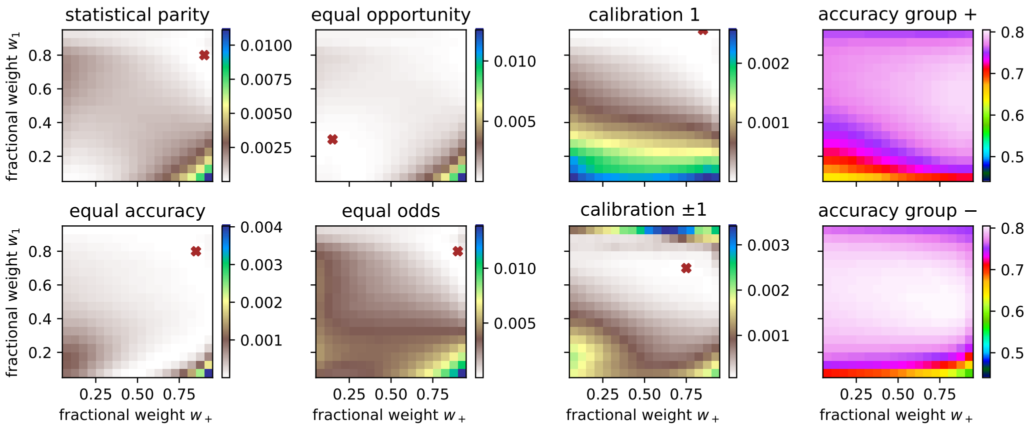

Figure F.4: **Mitigation using re-weighting on real data.**