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

Stefano Sarao Mannelli [*]
Data Science and AI, Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
and School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa

Federica Gerace [†]
Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, 40126 Bologna (BO), Italy

Negar Rostamzadeh [‡]
Google Responsible AI, Montreal, Canada

Luca Saglietti [§]
Computing Sciences Departments, Bocconi University, Milan, Italy

(Received 26 March 2024; revised 8 January 2025; accepted 8 July 2025; published 7 August 2025)

Machine learning (ML) may be oblivious to human bias but it is not immune to its perpetuation. Marginalization 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 to clarify 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 characterize 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. Simplifying the nature of the problem to its minimal components, we can retrace and unpack typical unfairness behavior observed on real-world datasets. Finally, we focus on the effectiveness of bias mitigation strategies, first by considering a loss-reweighing scheme that allows for an implicit minimization of different unfairness metrics and a quantification of the incompatibilities between existing fairness criteria. Then, we propose a mitigation strategy based on a matched inference setting that entails the introduction of coupled learning models. Our theoretical analysis of this approach shows that the coupled strategy can strike superior fairness-accuracy trade-offs.

DOI: 10.1103/nlfl-35t6

I. INTRODUCTION

Machine Learning (ML) systems are actively being integrated into multiple aspects of our lives, making the question about their failure points of utmost importance. Recent studies [1,2] have shown that these systems may have a significant disparity in failure rates across the multiple subpopulations targeted in the application. ML systems appear to perpetuate discriminatory biases that align with those present in our society [3–6].

Bias could originate at many levels in the ML pipeline, from the problem definition to data collection, to the training and deployment of the ML algorithm [7]. Without minimizing the importance of the other factors, we will focus this study on data itself, which often represents a critical source of bias [8]. A dataset can inadvertently contain the record of a history of discriminatory behavior, tangled in complex dependencies which are hardly eradicated even when the explicit discriminatory attribute is removed. The root of the discrimination can indeed be hidden in the structural properties of the dataset, since different subpopulations are almost inevitably heterogeneously represented. Thus, an important open question is *when and how such heterogeneity can induce bias in ML systems.*

Disproportional numerical representation of the different subpopulations in a dataset is of course the most visible—but not only possible—form of representation heterogeneity. Learning with an unbalanced dataset, where some classes are underrepresented, has been shown to drastically bias the outcome of a classifier [9,10]. Furthermore, imbalances in the relative representation can become particularly problematic in the high-dimensional, feature-rich regime [11]. In this work,

[*]Contact author: s.saraomannelli@chalmers.se
[†]Contact author: federica.gerace@unibo.it
[‡]Contact author: nrostamzadeh@google.com
[§]Contact author: luca.saglietti@unibocconi.it

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by Bibsam.

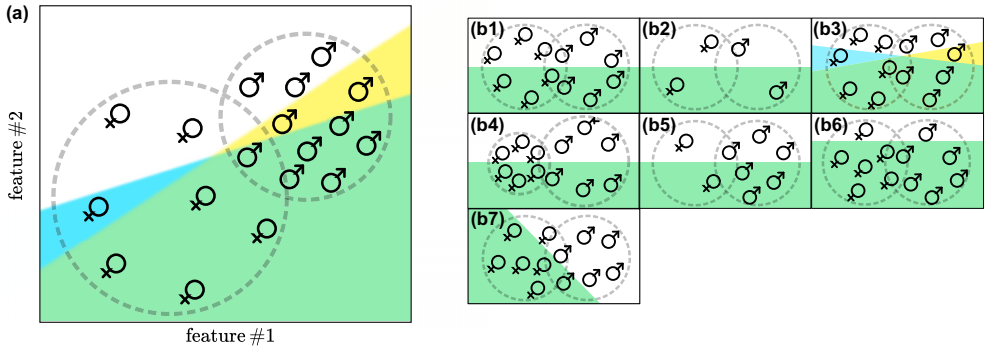


FIG. 1. The teacher-mixture (T-M) model can account for several types of data imbalance. (a) The T-M model is a generative model of high-dimensional structured data. Inputs are sampled from a combination of multivariate Gaussian distributions, with different centroids and covariances for each subpopulation in the dataset. The probability of sampling from each subpopulation can be tuned, giving rise to representation imbalance. In particular, the cartoon shows a larger relative representation for the male population (♂), which also has a smaller variance. The cyan and yellow shaded regions (green in their intersection) denote the decision boundaries of the labeling rules for the different data subpopulations, which in principle can be misaligned. Panel (b) exemplifies how manipulating the parameters of the T-M model can alter the data distribution: (b1) represents the *balanced condition* with equally represented, distributed and labeled samples; (b2) shows *scarcity* of data points in both clusters; (b3) displays an example of *rule misalignment*; (b4) shows different subpopulation *variances*; (b5) shows *relative representation* imbalance; (b6) represents the case of *unbalanced labels*; (b7) shows a case of positive *group-label correlation*.

however, we aim at identifying the *many other geometrical properties of data that can systematically lead to biased trained models.*

ML bias can be prevented or removed by implementing targeted heuristics in the training pipeline. A vast literature focuses on the study of bias mitigation methods in the context of real-world data, either by revising the data sampling step or by adjusting the optimization objective. Several methods have been shown to be effective in correcting for class imbalances in standard classification settings, including oversampling [12], undersampling [13], and reweighing strategies. In the general framework, the class label and subpopulation membership do not necessarily overlap but some of these ideas can be adapted to allow for bias mitigation [14,15]. Despite many empirical successes, a large gap remains in the theoretical understanding of bias-induction mechanisms and how to counteract them. The introduction of a *controlled minimal setting*, where these phenomena can be characterized exactly, could allow for a better theoretical grasp of these nuanced interactions.

In this work, we aim to address this theory gap by introducing the *teacher-mixture* (T-M) model, an exactly solvable generative model producing high-dimensional correlated data. This model offers a controlled setting where data imbalances and the emergence of bias become more transparent and can be better understood, allowing also for the design of theoretically grounded and effective solutions. The model is designed to capture common observations about the data structure of real datasets, with a particular focus on the coexistence of nontrivial correlations, both among inputs and between inputs and labels, induced by the presence of a subpopulation structure. Surprisingly, the few ingredients encoded in the model are capable of generating a rich and realistic ML bias phenomenology.

The rest of the work is structured as follows: in Sec. II we describe the T-M model and derive an analytical characterization of the typical performance of solutions in the high-dimensional limit. Section III examines the different sources of bias [shown in Fig. 1(b)] and their role in the bias-induction mechanism in a subpopulation-agnostic shallow network. This leads also to the identification of a *positive transfer* effect among the subpopulations within the dataset: despite their distinct characteristics, which make it tempting to split the dataset and use different classifiers, the shared underlying features can be leveraged to enhance the performance of a single classifier on both groups. Finally in Sec. IV, we focus on the problem of mitigating bias when the membership information is accessible. We theoretically analyze the effects of a sample reweighing mitigation strategy, highlighting the trade-offs between different definitions of fairness. We also propose and analyze a model-matched mitigation strategy, where two coupled networks are jointly trained, allowing for specialization on different subpopulations as well as transfer of valuable cross-population information.

II. MODELING DATA IMBALANCE

Drug testing provides a historically significant example of the potential consequences of unchecked data imbalance: substantial evidence [8,16,17] shows that the scarcity of data points corresponding to female individuals in drug-efficiency studies resulted in a larger number of side effects in their group. This historical data gap has often been justified on the basis of a "simplification" criterion: due to the inherent variance of the female subpopulation (caused, e.g., by fluctuating hormonal levels), their inclusion in medical trials can introduce complex interactions that are instead absent in the "standard" male subpopulation. However, ignoring biological sex as a discriminative factor in the analysis can induce serious adverse effects on the female subpopulation, ranging from over-dosage to ineffectiveness of treatment.

The T-M model is designed to allow a theoretical characterization of the impact of such data imbalances on the inference process (e.g., determining a discriminative rule for administering the drug to the patients). While retaining analytical tractability, the T-M model retraces the main features of real

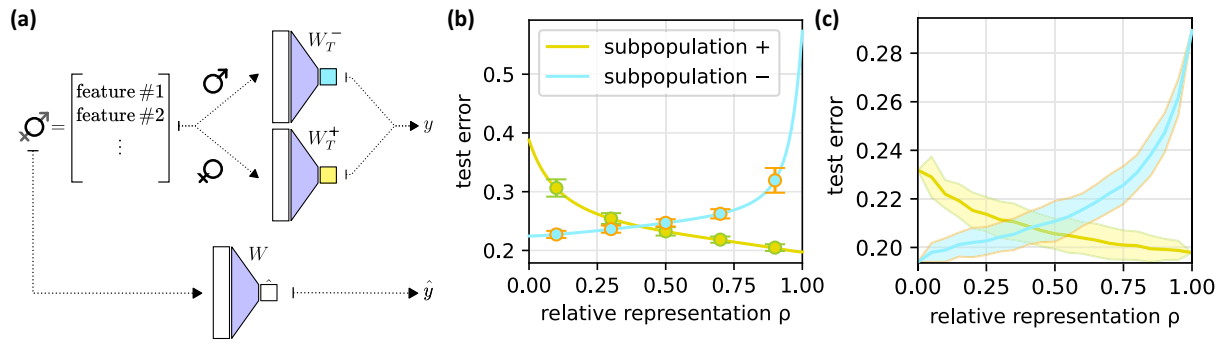


FIG. 2. Training on T-M model and comparison between error on synthetic and real data. (a) Given a vector of input features and a group membership (male/female), the ground-truth label is assigned by the associated one-layer teacher network (represented by one of the vectors $W_T^{\pm}$). The decision boundaries are demarked in blue and yellow [while their intersection is colored in green)]. The labeling rules can be aligned, i.e., the decision rule does not depend on the group membership, or misaligned as in panel (a). A one-layer student network is given inputs $\boldsymbol{x}^\mu$ and labels y^μ, and trained to produce the correct outputs $\hat{y}$ via gradient descent on the loss $\ell(\hat{y}, y)$. Panel (b) shows the test performance (on the two subpopulations) for a student network trained on mixed data instances with variable relative representations. Unsurprisingly, when one subpopulation is largely predominant in the dataset, the classifier becomes biased to have higher accuracy on it. The plot shows the match between the analytic curves (solid lines) and numerical simulations on the synthetic framework (dots). Panel (c) contains a similar experiment, but with data from the "CelebA" dataset [19]. Details in the Appendix B.

data with multiple coexisting subpopulations and allows for a richer phenomenology than previously analyzed data models. In Fig. 1(a), we sketch a two-dimensional cartoon of the T-M data distribution, framed in the context of drug testing.

The T-M combines aspects of two common modeling frameworks for supervised learning, namely the Gaussian-mixture (GM) and the teacher-student (TS) setups [18]. The GM is a simple model of clustered input data, where each data point is sampled from one out of a narrow set of high-dimensional Gaussian distributions. Instead, the T-M inherits from the TS setup a simple model of input-label correlation, where the ground-truth labels are produced by a realizable "teacher" rule, to be inferred by the trained model during the learning process. For simplicity, in this study, we will only consider linear labeling rules. In the T-M, however, we allow for the existence of group-specific rules: at inference, the model will have to strike a compromise between them. Many different factors, parametrically controlled in the T-M, can generate bias in a classifier, T-M allows one to explore different realizations of the problem as shown in Fig. 1(b).

In the sketch in Fig. 1(a), the female and male subpopulations are represented as partially overlapping data clouds, with different variances and group-dependent offsets (the two features in the sketch could represent some combination of clinical values recorded during the trial). Note that the female population is numerically under-represented, as in the above-described real-world scenarios. The shaded areas represent the true regions of the effectiveness of the tested drug for female/male subjects (cyan/yellow shades). As depicted in Fig. 2(a), the goal of the inference model is to infer a decision boundary for the administration of the drug based on the observations of its effectiveness on the test subjects. While the vast majority of the subjects would be identically classified according to the two different labeling rules (green region), some false positives could occur if the inference only accounts for a single subpopulation.

Figure 2 shows a representation of student trained on the T-M model. If no explicit bias mitigation strategy is employed, then a heterogeneous representation of the two subgroups will inevitably lead to a biased classifier. In Figs. 2(b) and 2(c), we show that the classification accuracy on the two subpopulations, as the fraction of data points belonging to each group (the relative representation) is varied, is biased in favor of the majority group.

For simplicity, the results discussed in this paper will focus on the case of two groups, but the analysis could be extended to multiple subpopulations.

Formal definition

We consider a synthetic dataset of n samples $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu, c^\mu\}_{\mu=1}^n$, with $\mathbf{x}^\mu \in \mathbb{R}^d$, $y^\mu, c^\mu \in \{1, -1\}$. We define the $\mathcal{O}(1)$ ratio $\alpha = n/d$ and we refer to it as the dataset size parameter. Each input vector is independent and identically distributed (i.i.d.), sampled from a Rademacher-Gaussians distribution with variances Δ_c, $\mathbf{x}|c \sim \mathcal{N}(c\boldsymbol{v}/\sqrt{d}, \Delta_c \mathbb{I}^{d\times d})$, where the random variable $c \sim \mathrm{Rad}_\rho$ denotes the subgroup membership. Notice that ρ $(1-\rho)$ controls the relative representation of subgroup $+$ (subgroup $-$, respectively) in the dataset. 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 [20,21], e.g., as in the case of CelebA and MEPS shown in the Appendix B. The ground-truth labels, instead, are provided by two Gaussian teacher vectors, namely $\mathbf{W}_T^+$ and $\mathbf{W}_T^-$, with respective bias terms b_T^+ and b_T^-, normalized to the d-dimensional sphere $\frac{1}{d}\mathbb{E}[\|\mathbf{W}_T^c\|^2] = 1$ and with mutual overlap $\frac{1}{d}\mathbb{E}[\mathbf{W}_T^+ \cdot \mathbf{W}_T^-] = q_T$. Each teacher produces labels for the inputs with the corresponding group-membership, namely $y^\mu = \mathrm{sign}(\mathbf{W}_T^{c^\mu} \cdot \mathbf{x}^\mu/\sqrt{d} + b_T^{c^\mu})$, with $c^\mu \in \{+, -\}$. The teacher bias terms are included in the model to control the fraction of positive and negative samples within the two

subpopulations. Overall, the geometric picture of the data distribution [sketched in Fig. 1(a)] is summarized by three sufficient statics, $m_T^c = \frac{1}{d}$, $\boldsymbol{W}_T^c \cdot \boldsymbol{v}$, and q_T, that respectively quantify the alignment of the teacher labeling rules with respect to the shift vector, controlling the group-label correlation, and the alignment 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 $\boldsymbol{W}$, with bias term b, producing outputs $\hat{y}^\mu = \text{sign}(\mathbf{W} \cdot \mathbf{x}^\mu/\sqrt{d} + b)$, and trained via empirical risk minimization (ERM) with loss:

$$\mathcal{L}(\boldsymbol{W}, b) = \sum_{\mu \in \mathcal{D}} \ell(\boldsymbol{W}, b; \boldsymbol{x}^\mu, y^\mu) + \frac{\lambda ||\boldsymbol{W}||_2^2}{2}, \tag{1}$$

where ℓ is assumed to be convex in student's parameters and λ is an external parameter that regulates the intensity of the L_2 regularization.

Given this framework, we derive a theoretical characterization of the training performance of this learning model and consider the possible implications from an ML fairness perspective. In particular, we aim to study the role of data geometry and cardinality in the training of a fair classifier. To quantify the level of bias in the predictions of the trained model, we need to choose a metric of fairness. We will employ *disparate impact* (DI) [22], an ML analog of the 80% rule [23], which allows a simple assessment of the over-specialization of the classifier on one of the subpopulations. In our framework, we characterize bias against subpopulation $+$ using the following definition of

$$\text{DI} = \frac{p(\hat{y} = y|+)}{p(\hat{y} = y|-)}, \tag{2}$$

evaluating the ratio between test accuracy in subpopulation $+$ and subpopulation $-$. Note that how to measure bias is itself an active line of research, and the DI alone cannot return a full picture of the unfairness. In Sec. IV, we compare these results with those obtained with other metrics. Notice that the T-M model allows one to parametrically move from a model-mismatched scenario ($q_T < 1$) where the rule to be inferred is not in the function space of learnable rules, to a model-matched scenario ($q_T = 1$) where the rule is actually learnable but, as we will discuss further in Sec. III, the model may systematically fail to identify it. We will discuss in detail when these failure modes occur and why.

Finally, the T-M model has, at the same time, the advantage of being simple, allowing a better understanding of the many facets of ML bias, and the disadvantage of being simple, since some modeling 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. However, this modeling approach continues a long tradition of research in statistical physics [24], which has shown that theoretical insights gained in prototypical settings can often be helpful in disentangling and interpreting the complexity of real-world behavior.

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

(i) the *relative representation*, $\rho = n_+/(n_+ + n_-)$, with n_c the number of points in group c, $c \in \{+, -\}$;

(ii) the *group variance*, Δ_c, determining the width of the clusters;

(iii) the *group label frequencies*, controlled through the bias terms b_T^c;

(iv) the *group-label correlation*, m_T^c;

(v) the *intergroup 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;

(vi) the *dataset size*, α, representing the ratio between the number of inputs and the input dimension.

Theoretical analysis in high-dimensions

In principle, solving Eq. (1) requires finding the minimizer of a complex nonlinear, high-dimensional, quenched random function. However, statistical physics [25] 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 behavior 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 sufficient statistics, $\Theta = \{Q = \frac{1}{d}\boldsymbol{W} \cdot \boldsymbol{W}, m = \frac{1}{d}\boldsymbol{W} \cdot \boldsymbol{v}, R_\pm = \frac{1}{d}\boldsymbol{W} \cdot \boldsymbol{W}_T^\pm, \delta q, b\}$, respectively representing the typical norm of the trained estimator, its magnetization in the direction of the cluster centers, its alignment with the two teacher vectors of the T-M, the rescaled variance of the self-overlap (details in Appendix A), and the student bias term. There are some technical quantities that emerge from the application of replica theory and are described in detail in Appendix A, here we give a brief overview. The parameter δq among the sufficient statistics represents the leading term in the low-temperature expansion of the self-overlap. Indeed following the classical replica method, the minimization problem is obtained studying a relaxed problem with a temperature parameter controlling the degree of relaxation, and then taking the limit of no-relaxation (temperature to zero). Furthermore, we will need the conjugated variables $\hat{\Theta} = \{\hat{Q}, \hat{m}, \hat{R}_\pm, \delta\hat{q}\}$ that play the role of Lagrange multipliers and enforce the order parameters in Θ (with the exception of b). Finally: The results below summarize the main findings of the replica analysis, while all the technical details are reported in Appendix A.

Analytical result 1. Given a specific setup of the T-M model, in the high dimensional limit when n, $d \to \infty$ at a fixed ratio $\alpha = n/d$, the scalar descriptors $\Theta = \{Q, m, R_\pm, \delta q, b\}$ of the vector $\boldsymbol{W}$ obtained by empirical risk minimization Eq. (1) with a generic convex loss ℓ, and their conjugated variables $\hat{\Theta} = \{\hat{Q}, \hat{m}, \hat{R}_\pm, \delta\hat{q}\}$, converge to deterministic quantities given by the unique fixed point of the system: $Q = -2\frac{\partial s(\hat{\Theta};\lambda)}{\partial \delta\hat{q}}$, $m = \frac{\partial s(\hat{\Theta};\lambda)}{\partial \hat{m}}$,

$R_\pm = \frac{\partial s(\hat{\Theta};\lambda)}{\partial \hat{R}_\pm}$, $\delta q = 2\frac{\partial s(\hat{\Theta};\lambda)}{\partial \hat{Q}}$, $\hat{Q} = 2\alpha\frac{\partial e(\Theta;\Delta)}{\partial \delta q}$, $\hat{m} = \alpha\frac{\partial e(\Theta;\Delta)}{\partial m}$, $\hat{R}_\pm = \alpha\frac{\partial e(\Theta;\Delta_\pm)}{\partial R_\pm}$, $\delta\hat{q} = 2\alpha\frac{\partial e(\Theta;\Delta)}{\partial Q}$, with

$$s(\hat{\Theta};\lambda) = \frac{1}{2(\delta\hat{q}+\lambda)}\Bigg[\hat{Q} + \Bigg(\hat{m} + \sum_{c\in\{\pm\}} m_T^c \hat{R}_c\Bigg)^2 + \sum_{c\in\{\pm\}}\left(1-\left(m_T^c\right)^2\right)\hat{R}_c^2 + 2\Bigg(q_T - \prod_{c\in\{\pm\}} m_T^c\Bigg)\prod_{c\in\{\pm\}}\hat{R}_c\Bigg], \tag{3}$$

$$e(\Theta;\Delta) = \mathbb{E}_c\Bigg[\mathbb{E}_z \sum_{y=\pm1} v(y,c,\Theta) \times H\Bigg(-y\frac{\sqrt{Q}\left(c\,m_T^c + b_T^c\right) + \sqrt{\Delta_c}R_c z}{\sqrt{\Delta_c\left(Q-R_c^2\right)}}\Bigg)\Bigg], \tag{4}$$

where $c \in \{+,-\} \sim \text{Bernoulli}(\rho)$, $z \sim \mathcal{N}(0,1)$, $H(\cdot) = \frac{1}{2}\text{erfc}(\cdot/\sqrt{2})$ is the Gaussian tail function. The function $v(y,c,\Theta)$, appearing in Eq. (4), depends parametrically on the scalar descriptors and entails a one-dimensional optimization problem: $v(y,c,\Theta) = \max_w[-\frac{w^2}{2} - \ell(y, \sqrt{\Delta_c\delta q}w + \sqrt{\Delta_c Q}z + c\,m + b))]$. The student bias term b implicitly solves the equation $\partial_b e(\Theta;\Delta) = 0$. Equations (3) and (4) represent the so-called entropic and energetic contributions appearing in the quenched free entropy of the system (details in Appendix B).

The yielded fixed point values for the scalar descriptors, Θ, can be used to obtain deterministic predictions for common model evaluation metrics, such as the *confusion matrix* or the *generalization error*, in high-dimensional realizations of the system.

The presented result was obtained through the nonrigorous replica method from statistical physics [18,25,26]. The derivation details are deferred to the Appendix A. We remark that, in convex settings, the replica method was rigorously proven to yield exact results in a range of different model settings. In particular, a lengthy but straightforward generalization of the proofs presented in Refs. [20,27,28] could be derived for the T-M case, but this is out of the scope of the present work. In this manuscript, we verify the validity of our replica theory by comparison with numerical simulations, as shown, e.g., in Fig. 2(b).

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 subpopulation c with true label y, are given by

$$p(\hat{y}\,|\,y;c) = \mathbb{E}_z\Bigg[\text{Heav}\left(y\left(\sqrt{\Delta_c}z + c\,m_T^c + b_T^c\right)\right) \times H\Bigg(-\hat{y}\frac{(c\,m+b) + \sqrt{\Delta_c}R_c z}{\sqrt{\Delta_c\left(Q-R_c^2\right)}}\Bigg)\Bigg], \tag{5}$$

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

This second result yields a fully deterministic estimate of the accuracy of the trained model on the different data subpopulations. These scores will be used in the following sections to investigate the possible presence of bias in the classification output of the model. In particular, they will be useful to estimate numerator and denominator of the DI, Eq. (2). While in the following sections we will specialize on cross-entropy loss, notice that the results apply to any convex loss function ℓ. Finally, notice that the results 1 and 2 allow for an extremely efficient and exact evaluation of the learning performance in the T-M, remapping the original high-dimensional optimization problem onto a system of deterministic scalar equations that can be easily solved by recursion.

III. INVESTIGATING THE SOURCES OF BIAS

We will use the analytical expression for the generalization performance reported in Eq. (5) to systematically investigate the effect of the sources of bias identified in Remark 1, which potentially mine the design of a fair classifier. In this section, we assume that the classifier is agnostic to the subgroup membership and effectively observes a dataset $\mathcal{D} = \{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^n$ to quantify the bias generated without any form of mitigation. We will reintroduce the subgroup membership c among the observed data entries in the discussion on mitigation strategies in Sec. IV. We specialize on cross-entropy loss and perform three separate experiments to summarize some distinctive features of the fairness behavior in the T-M: namely, the impact of the correlation between the labeling rules and the group structure, the interplay between relative representation and group variance, and the different accuracy trade-offs between the subpopulations at different dataset sizes. The parameters of the experiments, if not specified in the caption, are detailed in the Appendix A 1.

A. Group-label correlation

In Fig. 3(a), we consider a scenario where the labeling rules for the two groups are not perfectly aligned, i.e., $\boldsymbol{W}_T^+ \neq \boldsymbol{W}_T^-$ (and/or $b_T^+ \neq b_T^-$). 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 generalization for both subpopulations at the same time. For simplicity, we set an equal correlation between the two teacher vectors and the shift vector, $m_T^+ = m_T^- > 0$, and isolate the role of rule similarity q_T. The upper-left 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 exceeds the 80% rule when the labeling rules are misaligned and the group sizes are numerically unbalanced (small q_T and ρ). Indeed, in the cut displayed in the upper-right panel, by lowering the group-label correlation $m_T^\pm$ the gap between

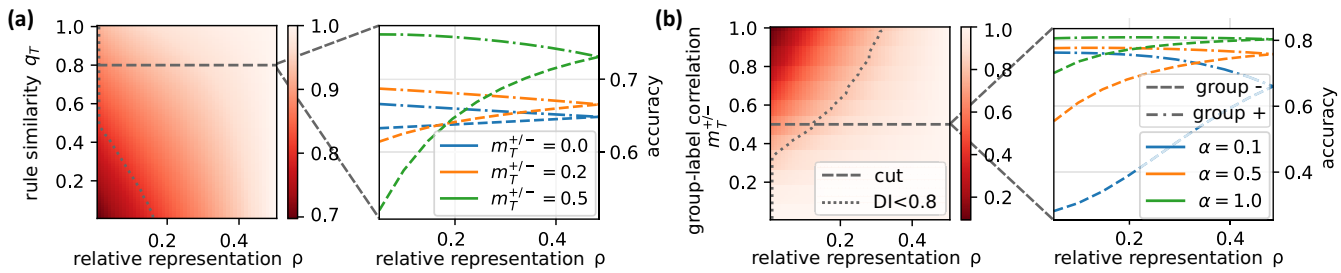


FIG. 3. Simple geometrical properties cause the emergence of bias. Each point in the left diagrams shows, for different values of the model parameters, the Disparate Impact (DI) of the trained model (darker colors represent stronger biases). In particular, in the left diagrams, on the x axis we vary the relative representation ρ, while on the y axis we explore possible values of the rule similarity q_T for panel (a) and the group-label correlation $m_T^{\pm}$ for panel (b).The corresponding figures on the right show the values of the accuracy for the two subpopulations in correspondence of the cut represented by the dashed line on the left.

the measured accuracies on the two subpopulations becomes smaller. However:

Remark 2. Even when $q_T = 1$ and the task is solvable (i.e., the model is not mismatched and the classifier can learn the input-output mapping), the trained model can still be biased.

This is shown in Fig. 3(b), where a large high-bias region (DI<80%) exists. In particular, the lower-left panel shows the cause of this effect in the presence of a nonzero group-label correlation $m_T^{\pm}$, and in the lower-right panel we see how this effect is more pronounced in the data-scarce regime. In all four panels, as ρ reaches 0.5, the two subpopulations become equally represented and the classifier achieves the same accuracy for both.

B. Bias and variance

In Fig. 4, 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.

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 left panel indicate a higher accuracy for the smaller subpopulation even if the dataset only contains 10% of samples belonging to it. This exemplifies the fact that a very focused distribution (low $\Delta_{\pm}$) actually requires less samples. The right 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 vice versa) one induces a bias in the classification.

C. Positive transfer

If mixing different subpopulations in the same dataset can induce unfair behavior, then one could consider splitting the data and train independent models. In Fig. 5, we show that a *positive transfer effect* [29] can yet be traced between the two groups when the rules are sufficiently similar. This means that the accuracy on the under-represented group is enhanced when information is shared across the two subpopulations.

Remark 4. The performance on the smaller subpopulation tends to further deteriorate if the dataset is split according to the subgroup structure.

To clarify this point, in the left plot of Fig. 5 we show the DI as a function of the dataset size α, for several values of the rule similarity q_T and at fixed $\rho = 0.10$. In the center and right plots in Fig. 5, we also display the gain in accuracy on each subpopulation when the model is trained on the full dataset, comparing with a baseline classifier (black lines) trained only

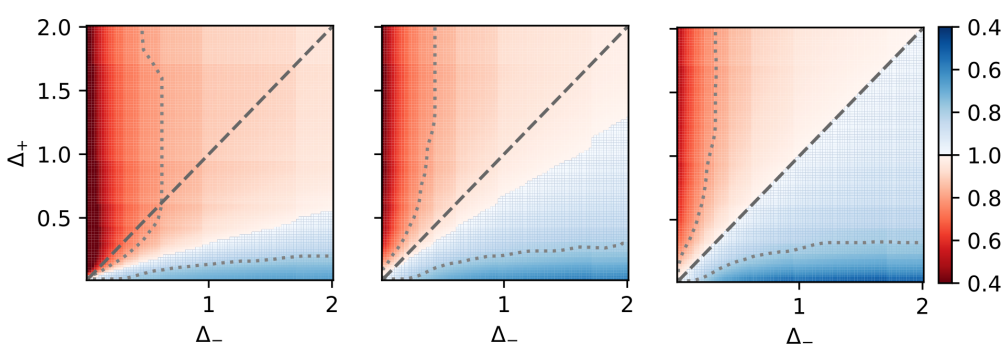


FIG. 4. Emergence of bias even in balanced datasets. We show the disparate impact as the distribution of the two subpopulations is changed by altering their variances (Δ_+ and Δ_-). The diagonal line gives the configurations where the two subpopulations have the same variance. The two figures consider 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 colors to quantify the disparate bias against subpopulation $+$ and $-$, respectively.

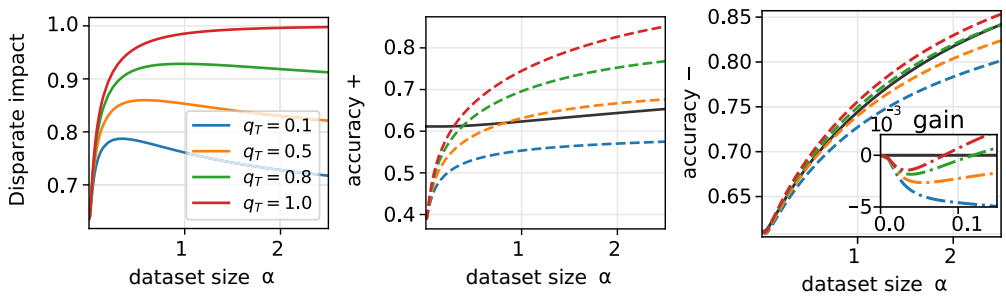


FIG. 5. Performance benefits for both subpopulations under shared training. With 10% of the data points in subpopulation + ($\rho = 0.1$), we compare the performance with different levels of rule similarity (q_T) as the size of the dataset is increased, showing the disparate impact in the left figure and the individual accuracies in central and right ones. In central and right figures, the baselines—plotted in black—show the accuracies attained when the model is trained only on the corresponding group data. The inset of the rightmost figure highlights the differences in accuracy in the small dataset regime. When the rules are sufficiently aligned, joint training on both groups will induce a better accuracy on the smaller subpopulation provided α is not too small. Moreover, at intermediate values of α also the larger group can benefit from the information transfer.

on the respective data subsets (+ in the central panel, − in the lower panel). These two plots elucidate the positive transfer effect: for sufficiently similar rules (large q_T), both populations can benefit from shared training at intermediate dataset sizes. If the dataset is too small (low α regime), then the lack of data combined with a high variance in the input distribution can induce over-fitting, with a larger drop in performance for the smaller group. However, as the dataset size becomes sufficiently large, the positive transfer effect is eventually lost for the large subpopulation (large α regime).

Connecting the results of this section to the drug testing examples, after observing that the distribution of side effects in the female population presented a larger variance, the solution was simply to collect more data of female subjects. Moreover, the results shown in Fig. 4 indicate the need for a higher relative representation of female subjects in the dataset to achieve an unbiased classifier. While implementing this solution might have introduced higher variance in the results due to the intrinsic high-variability of the data, it would have significantly reduced the risk of administering drugs with limited testing on half of the population.

IV. MITIGATION STRATEGIES

To assess or ensure the fairness of a ML model on a given data distribution, a plethora of different fairness criteria have been designed [30,31]. In convex settings, any of these criteria can be separately enforced via a hard constraint during the optimization process [32–34]. However, it was proved that some criteria are completely incompatible and cannot be exactly achieved simultaneously [35–37]. In the same spirit of Ref. [30], we drop the hard constraint and instead quantify exactly how far a given trained model is from meeting the criteria. Each criterion requires the probability of obtaining a specific classification outcome E to be the same across the subpopulations. For example, according to the definition of *equal opportunity* (Table I), the *true positive* rate $P(E = (\hat{Y} = 1 \mid Y = 1))$ should not depend on the group-membership C. A natural measure of the observed dependence between E and C is given by the mutual information (MI):

$$I(E;C) = D_{KL}(\mathbb{P}[E,C] \mid \mathbb{P}[E]\mathbb{P}[C]) = \mathbb{E}\log\frac{\mathbb{P}[E,C]}{\mathbb{P}[E]\mathbb{P}[C]}. \tag{6}$$

The fairness condition is exactly verified only when the joint distribution factorizes, i.e., $\mathbb{P}[E,C] = \mathbb{P}[E]\mathbb{P}[C]$, and the mutual information goes to zero. Table I provides some other examples of classification events E, for some well-established fairness criteria. Note that some criteria might not be sensible in specific settings (e.g., *Statistical Parity* is unlikely to be guaranteed in a drug-testing scenario).

In the following, we consider two simple bias mitigation strategies that can be analyzed within our analytical framework. The required generalizations of the results are detailed in the Appendix A. First, we study the debiasing effect of a sample reweighing strategy where the relevance of each sample is varied based on its label and group membership [38–40]. By adjusting the weights, one can indirectly minimize 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 multitask learning [41], that couples two architectures trained in parallel.

A. Loss reweighing

Recent literature shows that some fairness constraints cannot be satisfied simultaneously. ML systems are instead forced to accept trade-offs between them [35]. This sort of compromise is well-captured in the simple framework of the T-M model. The first three panels of Fig. 6(a) show accuracies and 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. Thus, each loss term in Eq. (1) is reweighed

TABLE I. List of Fairness Metrics. *Statistical Parity*: Equal fractions of each group should be treated as belonging to the positive class [35,42,43]. *Equal Opportunity*: Each group needs to achieve equal true positive rate [44]. *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 rates [22,45]. *Predicted Parity*. Given inputs that are classified by the model with label y, the fraction of input with true label y^* should be consistent across subpopulations. This gives two subcriteria: *predicted parity 1* requires the condition only for $y^* = 1$, while *predicted parity 10* requires the condition for both $y^* = 1$ and $y^* = -1$ [46].

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	$\mathbb{P}[\hat{Y} = 1 \vert C = c, Y = 1] = \mathbb{P}[\hat{Y} = 1 \vert Y = 1] \cap \mathbb{P}[\hat{Y} = 1 \vert C = c, Y = -1] = \mathbb{P}[\hat{Y} = 1 \vert Y = -1]\ \forall c$
Predicted parity	$\mathbb{P}[Y = 1 \vert C = +, \hat{Y} = y] = \mathbb{P}[Y = 1 \vert C = -, \hat{Y} = y] = \mathbb{P}[Y = 1 \vert \hat{Y} = y]\ \forall y$

as

$$\mathcal{W}(c, y) = \begin{cases} w_1 w_+ & \text{if} \quad c = +,\ y = 1, \\ w_1(1 - w_+) & \text{if} \quad c = -,\ y = 1, \\ (1 - w_1) w_+ & \text{if} \quad c = +,\ y = -1, \\ (1 - w_1)(1 - w_+) & \text{if} \quad c = -,\ y = -1. \end{cases} \tag{7}$$

By changing these relative weights one can force the model to pay more attention to some types of errors and reestablish a balance between the accuracies on the two subpopulations. The goal is to identify a classifier that achieves high accuracy (lower panels) while minimizing the MI for different fairness metrics. Notably, given a weight w_1, these minima occur for different values of the weight w_+. Only $w_1 = 0.1$ seems to have a value of w_+ close to several minima of the MI, but this point correspond to a sharp decrease in accuracy in both subpopulations, thus fairness is achieved but at the expense of accuracy. These results are in agreement with rigorous results in the literature [37], but also show how the incompatibilities between the different constraints extend to regimes where the fairness criteria are not exactly satisfied.

B. Coupled networks

The emergence of classification bias in the T-M traces back to the clear mismatch between the generative model of data

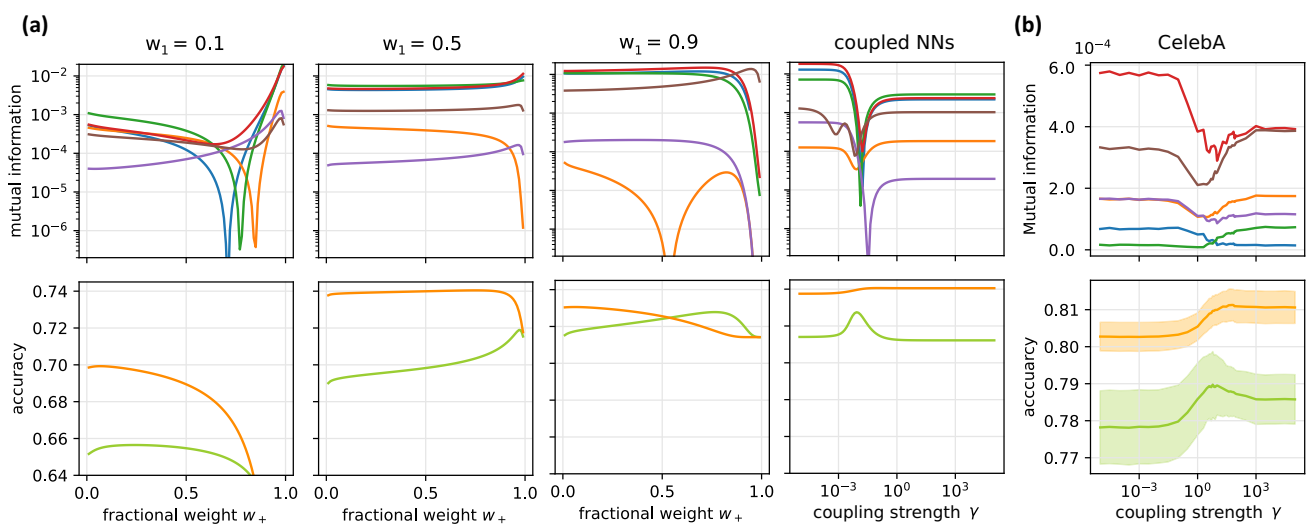


FIG. 6. Fairness-accuracy trade-off with reweighing and coupled architecture. (a) The figures show the effect of re-weighting and coupled architectures debiasing methods in a instance of the T-M model. The lowers figures shows the accuracy for subpopulation + and subpopulation − and the upper figures show the mutual information for the several fairness metrics defined in Table I, namely statistical parity, equal opportunities, equal accuracy, equal odds, predicted parity 1, and predicted parity 10. The goal of the algorithm is to identify regions with high accuracy (lower figures) and low mutual information (higher figures) for all metrics: this would imply that fairness is approximately achieved under all the criteria. The first three group of figures refer to the reweighing strategy, forcing higher relevance for a certain label in each panel ($w_1 = 0.1, 0.5, 0.9$) and the relative importance of a given subpopulation (parameter w_+) on the x axis. The last panels instead refer to the proposed coupled networks strategy and the x axis represent the strength of the coupling γ. The figures clearly show that our strategy achieves a higher accuracy in both subpopulations while preserving a higher level of fairness. Interestingly the minimum of the mutual information roughly correspond to the same parameter of the coupling strength, contrarily to what observed in the reweighing strategy. (b) The two panels, show an example from the CelebA dataset splitting and classifying according to the attributes "Wearing_Lipstick" and "Wavy_Hair," respectively, more details are provided in the Appendixes B and B 1. The observations made for the synthetic model applies also in this real-world case.

and the learning model. To move toward a matched inference setting, we need to enhance the learning model to account for the presence of multiple subpopulations and labeling rules. This inspires a mitigation strategy that we call *coupled neural networks*, consisting in the simultaneous training of multiple neural networks, each one seeing a different subset of the data associated with a different subpopulation. This idea is represented by the following modified loss:

$$\mathcal{L}_{\rm cnn}(\boldsymbol{w}) = \sum_{c\in\pm}\sum_{\mu\in\mathcal{D}^c} \ell(\boldsymbol{W}_c, b_c; \boldsymbol{x}^\mu, y^\mu) + \frac{\lambda}{2}\left(||\boldsymbol{W}_+||_2^2 + ||\boldsymbol{W}_-||_2^2\right) + \frac{\gamma}{2}||\boldsymbol{W}_+ - \boldsymbol{W}_-||_2^2, \tag{8}$$

where $\boldsymbol{W}_\pm$ are the weights of the two networks, $b_\pm$ their associated bias terms, and $\hat{y}_\pm^\mu$ are their respective estimation of label y^μ. The networks exchange information through the elastic penalty γ that mutually attracts them, and the intensity of this elastic interaction is obtained by cross-validation. In the extreme case of orthogonal teachers $q_T = 0$ the optimal performance will be achieved for zero coupling $\gamma = 0$, while in the case of identical teachers $q_T = 1$ the optimal coupling will tend to infinity. This approach shares some ideas with other methods present in the literature: Refs. [47,48] add an elastic penalty term to the loss to bias the training trajectory, while Ref. [49] proposes to combine estimations from different Bayesian classifiers. However, note that these prior approaches were tailored for different learning protocols or problem domains, and could not be applied in the problem setting considered in this paper. Other similar optimization strategies include simultaneous linear regression [50] and multiple factor analysis, but to our knowledge, these approaches have not yet been applied in the bias mitigation context.

Remark 5. The *coupled neural networks* method allows for higher expressivity and specialization on the various subpopulations, while also encouraging positive transfer between similarly labeled subpopulations, leading to better fairness-accuracy trade-offs.

The upper rightmost plots of Fig. 6(a), displaying the behavior 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 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 minimizes the gap in terms of test accuracy between the two subpopulations, as shown in the lower plot, without hindering the performance on the larger group. Notice that this result does not contradict the impossibility theorem [37] which states that statistical parity, equal odds, and predicted parity cannot be satisfied altogether. In fact, our result only concerns soft minimization of each fairness metrics. The result is in agreement with Ref. [51] whose results show that the trade-off between fairness and accuracy vanishes when the true distribution of data is capture. Leveraging the universal approximation property of neural networks, the coupled networks method seems a promising direction for applications. In the panels of Fig. 6(b) we show promising preliminary results in the realistic dataset CelebA [52]. We stress that real data often presents more complex correlations than those modeled in the T-M, which may hinder the effectiveness of this strategy in unexpected ways.

The method of the coupled networks can be generalized to an arbitrary number of classes and subpopulations, and can be combined with standard clustering methods when the group membership label is not available. A future research direction will be to better investigate its range of applicability and, consequently, its limitations in real-world scenarios. In the Appendixes A and A 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.

V. DISCUSSION

The goal of this study was to design a generative model of high-dimensional correlated data that allows the study of the effect of data geometry in the bias induction mechanism, in isolation from real-world confounding factors. While a focus on each specific dataset might be required to ensure fairness in applications with high societal impact, we believe that the study of the ML bias phenomenology in controlled synthetic settings might allow more coherent advancements in the understanding and prevention of bias induction.

The T-M model captures nontrivial correlations among inputs and between inputs and labels, representing various imbalances appearing in real datasets when different subpopulations coexist in the sample. Surprisingly, with few modeling ingredients, the T-M can generate a rich and realistic ML bias phenomenology. We derive an analytical characterization of its performance in the high-dimensional limit, showing agreement with numerical simulations and producing realistic unfairness behavior. By isolating different sources of bias, we gain insights into situations where unfairness may persist despite apparent data balance, cautioning against relying solely on simple rebalancing techniques. We identify a positive transfer effect among diverse subpopulations, leveraging shared underlying features to enhance performance across groups. Additionally, we analyzed the trade-offs between different ways of quantifying the model fairness, focusing on a sample reweighing mitigation strategy that can be analytically characterized within our framework. We also proposed a theory-based mitigation strategy that effectively promotes fairness without compromising overall performance, as demonstrated in the T-M model. Instead of imposing an hard constraint on a desired fairness metric which would incur in the incompatibility theorem [37], the coupled networks strategy minimizes several fairness metrics simultaneously only approximately. Furthermore, the strategy seems to avoid the typical fairness-accuracy trade-off. This result is in agreement with the findings of Ref. [51].

Moving forward, our model is extremely simplified with the respect to real data and practical architectures. Future directions for our research include incorporating more complex elements into the data model considering model complex data structures, for instance assuming that data live in a low-dimensional manifold of the input space [53] or moving

away from Gaussian setting [54], and introducing the effect of feature dependencies (e.g., proxy variables) in the generated data. Another important but challenging address for further work is to move to the nonconvex optimization setting and to more complex model architectures, where at this time some of the analytic techniques employed in this work fail to generalize. However, recent works succeeded in addressing some of this limitations, in particular considering the multilabel classification problem [55] and more than a single layer [28]—despite still limited to random projection—these results could be included in future iterations of the work. Moreover, further explorations of the efficacy and limitations of the coupled networks strategy in the context of deep networks and more complex datasets is called for. By investigating its performance in deeper architectures and diverse real-world datasets, and connecting to the existing literature [56,57], we can assess the scalability and generalizability of this approach for addressing fairness concerns.

ACKNOWLEDGMENTS

The authors thank Giulia Bassignana and Andrew Saxe for many insightful discussions related to this work. We are also grateful to Sharat Chikkerur, Marc Mézard, and Stephen Pfohl for their valuable feedback on earlier drafts of the manuscript. This research was supported by the Wallenberg AI, Autonomous Systems, and Software Program (WASP) under Grant No. 219627/Z/19/Z, and in its early stages by the Gatsby Charitable Foundation (Grant No. GAT3755).

APPENDIX A: REPLICA ANALYSIS

We will directly present the most general setting for this calculation, where the learning model is composed of two linear classifiers ("students" in the following), coupled by an elastic penalty of intensity γ. This allows us to characterize the mitigation strategy proposed in this work, while the standard case with a single learning model can be obtained by setting $\gamma = 0$. Each student, denoted by the index $s = 1, 2$, is assumed to be trained on a fraction of the full dataset $\mathcal{D}_s$. Note that, in principle, the data split could not be aligned with the group structure of the dataset.

The loss function for the coupled learning model reads

$$\mathcal{L}(\boldsymbol{W}_1, \boldsymbol{W}_2) = \sum_{s=1,2} \sum_{\mu \in \mathcal{D}_s} \ell\left(\frac{\boldsymbol{W}_T^{c^\mu} \cdot \boldsymbol{x}^\mu}{\sqrt{d}} + b_T^{c^\mu}, \frac{\boldsymbol{W}_s \cdot \boldsymbol{x}^\mu}{\sqrt{d}} + b_s\right) + \sum_{s=1,2} \frac{\lambda}{2}\left(\sum_{i=1}^{d} W_{s,i}^2\right) - \frac{\gamma}{2}\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|^2, \tag{A1}$$

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

$$\ell(y, q) = -\mathrm{Heav}(y) \log \sigma(q) - (1 - \mathrm{Heav}(y)) \log (1 - \sigma(q)), \tag{A2}$$

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

1. Teacher partition function

In the T-M model, the label distribution is nontrivially 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 $\boldsymbol{W}_T^{\pm}$. Since we are allowed to fix a Gauge for one of these vectors (compatible with its distribution), we choose for simplicity $\boldsymbol{v} = \mathbf{1}$ to be a vector with all entries equal to 1 (still normalized on the sphere of radius d). We define the teacher partition function:

$$\begin{aligned} Z_T &= \int d\mu(\boldsymbol{W}_T^+, \boldsymbol{W}_T^-) \\ &= \int \prod_{c=\pm} \left[d\mu(\boldsymbol{W}_T^c)\delta\left(\left|\boldsymbol{W}_T^c\right|^2 - d\right)\delta\left(\boldsymbol{W}_T^c \cdot \mathbf{1} - d m_T^c\right)\right]\delta(\boldsymbol{W}_T^+ \cdot \boldsymbol{T}_- - d\, q_T), \end{aligned}$$

where the measures $\mu(\boldsymbol{T}_\pm)$ are in this case assumed to be factorized normal distributions. The Dirac's δ functions ensure that the geometrical disposition of the model vectors is the one defined by the chosen magnetizations $m_T^\pm$ and the overlap q_T, and that the vectors are normalized to the d-sphere.

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

$$\delta(x - ad) = \int \frac{d\hat{a}}{2\pi/d} e^{-i\hat{a}(\frac{x}{d} - a)}, \tag{A3}$$

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

$$Z_T = \int \prod_{c=\pm} \frac{d\hat{Q}_T^c}{2\pi/d} \int \prod_{c=\pm} \frac{d\hat{m}_T^c}{2\pi/d} \int \frac{d\hat{q}_T}{2\pi/d} e^{d\,\Phi_T(\{m_T^\pm, q_T\}, \{\hat{Q}_T^\pm, \hat{m}_T^\pm, \hat{q}_T\})}, \tag{A4}$$

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 can be factorized over them. In high dimensions, i.e., when $d \to \infty$, 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 to avoid dealing explicitly with imaginary quantities, and decompose $\Phi_T = g_{Ti} + g_{Ts}$:

$$g_{Ti} = -\left(\sum_c \hat{m}_T^c m_T^c + \sum_c \hat{Q}_T^c + \hat{q}_T q_T\right),$$
$$g_{Ts} = \log \int \mathcal{D}T_+ \int \mathcal{D}T_- \exp\left(\sum_c \hat{Q}_T^c T_c^2 + \sum_c \hat{m}_T^c T_c + \hat{q}_T T_+ T_-\right). \tag{A5}$$

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

$$g_{Ts} = \frac{(1-2\hat{Q}_T^-)(\hat{m}_T^+)^2 + (1-2\hat{Q}_T^+)(\hat{m}_T^-)^2 + 2\hat{q}_T \hat{m}_T^+ \hat{m}_T^-}{2\big((1-2\hat{Q}_T^-)(1-2\hat{Q}_T^+) - \hat{q}_T^2\big)} - \frac{1}{2}\log\big((1-2\hat{Q}_T^+)(1-2\hat{Q}_T^-) - \hat{q}_T^2\big).$$

Now, to complete the computation of the partition function Z_T, we have impose the saddle-point condition for Φ_T, which is realized when the entropy is extremized with respect to the fields we introduced. From the associated saddle-point equations one can find two useful identities:

$$1 - \left(m_T^c\right)^2 = \frac{\left(1-2\hat{Q}_T^{-c}\right)}{\left((1-2\hat{Q}_T^-)(1-2\hat{Q}_T^+) - \hat{q}_T^2\right)}, \tag{A6}$$
$$q_T - m_T^+ m_T^- = \frac{\hat{q}_T}{\left((1-2\hat{Q}_T^-)(1-2\hat{Q}_T^+) - \hat{q}_T^2\right)}. \tag{A7}$$

2. Free entropy of the learning model

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

$$\boldsymbol{W}_1^\star,\ \boldsymbol{W}_2^\star = \underset{\boldsymbol{W}_1,\boldsymbol{W}_2}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{W}_1,\ \boldsymbol{W}_2;\mathcal{D}), \tag{A8}$$

where $\mathcal{D}$ represents a realization of the data and $\mathcal{L}(\cdot)$ was defined in Eq. (A1). 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:

$$P(\boldsymbol{W}_1,\ \boldsymbol{W}_2;\mathcal{D}) = \frac{e^{-\beta\mathcal{L}(\boldsymbol{W}_1,\boldsymbol{W}_2;\mathcal{D})}}{Z_W}, \tag{A9}$$

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 (normalization 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 minimizer of the loss, representing the solution of the learning problem. In the asymptotic limit $d \to \infty$, the behavior of this model becomes predictable since the overwhelming majority of the possible dataset realizations (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 realization of the dataset so that the typical learning scenario can be captured.

Thus, we aim to compute the average free energy:

$$\Phi_W = \lim_{d\to\infty}\lim_{\beta}\frac{1}{\beta d}\langle \log Z_W(\boldsymbol{W}_1,\boldsymbol{W}_2;\mathcal{D}_1,\mathcal{D}_2)\rangle_{\mathcal{D}_1,\mathcal{D}_2}. \tag{A10}$$

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 $\lim_{r\to 0}(x^r - 1)/r = \log(x)$, provides a method to tackle this computation. One can replicate the partition function, introducing r independent copies of the original system. Each of them, however, sees the same realization of the data $\mathcal{D}$ (the "disorder" of the system, in the statistical physics terminology). When one takes the average over $\mathcal{D}$, the r 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 r to the real axis and computes the limit $\lim_{r\to 0}$, reestablishing the logarithm and the initial expression.

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

$$\Omega^r(\mathcal{D}) = \int \frac{d\mu(\boldsymbol{W}_T^+,\boldsymbol{W}_T^-)}{Z_T}\int\prod_{s,a}\left[db_s^a d\boldsymbol{W}_s^a e^{-\frac{\beta\gamma}{2}\|\boldsymbol{W}_1^a-\boldsymbol{W}_2^a\|^2}\prod_{\mu\in\mathcal{D}_s} e^{-\beta\,\ell\left(\frac{\boldsymbol{W}_T^{c^\mu}\cdot\boldsymbol{x}^\mu}{\sqrt{d}}+b_T^{c^\mu},\ \frac{\boldsymbol{W}_s^a\cdot\boldsymbol{x}^\mu}{\sqrt{d}}+b_s^a\right)}\right], \tag{A11}$$

where $s = 1, 2$ indexes the two coupled student models and $a = 1, ..., r$ 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. We can exploit δ functions 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:

$$1 = \int \prod_\mu du_\mu\, \delta\left(u_\mu - \frac{\boldsymbol{W}_T^{c^\mu} \cdot \boldsymbol{x}^\mu}{\sqrt{d}}\right) \int \prod_{a,s,\mu\in\mathcal{D}_s} d\lambda^a_\mu \delta\left(\lambda^a_\mu - \frac{\boldsymbol{W}_s^a \cdot \boldsymbol{x}^\mu}{\sqrt{d}}\right) \tag{A12}$$

$$= \int \prod_\mu \frac{du_\mu d\hat{u}_\mu}{2\pi} e^{i\hat{u}_\mu (u_\mu - \sum_{i=1}^d \frac{W_{T,i}^{c^\mu} x_i^\mu}{\sqrt{d}})} \int \prod_{a,s,\mu\in\mathcal{D}_s} \frac{d\lambda^a_\mu d\hat{\lambda}^a_\mu}{2\pi} e^{i\hat{\lambda}^a_\mu (\lambda^a_\mu - \sum_{i=1}^d \frac{W_{s,i}^a x_i^\mu}{\sqrt{d}})}. \tag{A13}$$

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

$$\mathbb{E}_{\boldsymbol{x}^\mu} e^{-i\sum_a \hat{\lambda}^\mu_a \sum_{i=1}^N \frac{W_{s^\mu,i}^a x_i^\mu}{\sqrt{d}} - i\hat{u}_\mu \sum_{i=1}^d \frac{W_{T,i}^{c^\mu} x_i^\mu}{\sqrt{d}}}$$

$$= \prod_{i=1}^N e^{-ic^\mu (\sum_a \hat{\lambda}^\mu_a \frac{W_{s^\mu,i}^a v_i}{d} + \hat{u}^\mu \frac{W_{T,i}^{c^\mu} v_i^\mu}{d})} \mathbb{E}_{z_i^\mu} e^{-i(\sum_a \hat{\lambda}^\mu_a \frac{W_{s^\mu,i}^a}{\sqrt{d}} + \hat{u}_\mu \frac{W_{T,i}^{c^\mu}}{\sqrt{d}}) z_i^\mu} \tag{A14}$$

$$= e^{-ic^\mu (\sum_a \hat{\lambda}^\mu_a \frac{\sum_i W_{s^\mu,i}^a}{d} + \hat{u}^\mu \frac{\sum_i W_{T,i}^{c^\mu}}{d}) - \frac{\Delta_{c^\mu}}{2} (\sum_{ab} \hat{\lambda}^\mu_a \hat{\lambda}^\mu_b \frac{\sum_i W_{s^\mu,i}^a W_{s^\mu,i}^b}{d} + 2\hat{u}^\mu \sum_a \hat{\lambda}^\mu_a \frac{\sum_i W_{s^\mu,i}^a W_{T,i}^{c^\mu}}{d} + (\hat{u}^\mu)^2 \frac{\sum_i (W_{T,i}^{c^\mu})^2}{d})}. \tag{A15}$$

To get Eq. (A15), we used the fact that the noise $\boldsymbol{z}^\mu$ is i.i.d. sampled from centered Gaussians of variance determined by the group, and explicitly used our Gauge choice $\boldsymbol{v} = \boldsymbol{1}$. In this expression, the relevant order parameters of the model appear, 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:

(i) $m_s^a = \frac{\boldsymbol{W}_s^a \cdot \boldsymbol{1}}{d}$, $m_T^c = \frac{\boldsymbol{W}_T^c \cdot \boldsymbol{1}}{d}$: magentizations in the direction of the + group center of the students and the teachers.

(ii) $q_s^{ab} = \frac{\sum_i w_{si}^a w_{si}^b}{d}$: self-overlap between different replicas of each student.

(iii) $R_{sc}^a = \frac{\sum_i W_{s,i}^a W_{T,i}^c}{d}$: overlap between student and teacher vectors.

(iv) $q_T^c = \frac{\sum_i T_{ci}^2}{d}$: norm of the teacher vectors (equal to 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

$$\Omega^d = \int \prod_{s,a} \frac{dm_s^a d\hat{m}_s^a}{2\pi/d} \int \prod_{sc,a} \frac{dR_{sc}^a d\hat{R}_{sc}^a}{2\pi/d} \int \prod_{s,ab} \frac{dq_s^{ab} d\hat{q}_s^{ab}}{2\pi/d} \int \prod_c db_c^a G_I^d G_S^d \prod_{sc} G_E(s,c)^{\alpha_{c,s} d}, \tag{A16}$$

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:

$$G_I = \exp\left(-\sum_{s,a} \hat{m}_s^a m_s^a - \sum_{s,ab} \hat{q}_s^{ab} q_s^{ab} - \sum_{sc,a} \hat{R}_{sc}^a R_{sc}^a\right), \tag{A17}$$

$$G_S = \int \prod_c \mathcal{D}T_c \exp\left(\sum_c \hat{Q}_T^c T_c^2 + \sum_c \hat{m}_T^c T_c + \hat{q}_T T_+ T_-\right) \int \prod_{s,a} d\mu(w_s^a) e^{-\beta\gamma (w_1^a - w_2^a)^2}$$

$$\times \exp\left(\sum_{s,a} \hat{m}_s^a w_s^a + \sum_{s,ab} \hat{q}_s^{ab} w_s^a w_s^b + \sum_{sc,a} \hat{R}_{sc}^a w_s^a T_c\right), \tag{A18}$$

$$G_E(s,c) = \int \frac{du d\hat{u}}{2\pi} e^{iu\hat{u}} \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a}\right) e^{-\frac{\Delta^c}{2} \sum_{ab} \hat{\lambda}_a \hat{\lambda}_b q_s^{ab} - \Delta^c \hat{u} \sum_a \hat{\lambda}_a R_{sc}^a - \frac{\Delta^c}{2} (\hat{u})^2} \prod_a e^{-\beta\, \ell(u + c m_T^c + b_T^c, \lambda^a + c m_s^a + b_s^a)}. \tag{A19}$$

The shorthand notation $\mathcal{D}x = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ is used to indicate a normal Gaussian measure. Note that, after the factorization in the G_S, the variables T_c and w_s^a denote a component of the vectors $\boldsymbol{W}_T^c$ and $\boldsymbol{W}_s^a$, respectively.

a. 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 optimization objective (A1), 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 r replicas of the students to be indistinguishable and the free entropy to be invariant under their permutation. Mathematically, the RS ansatz implies that

(i) $m_s^a = m_s$ for all $a = 1, ..., r$ (same for the conjugate),
(ii) $R_{sc}^a = R_{sc}$ for all $a = 1, ..., r$ (same for the conjugate),
(iii) $q_s^{ab} = q_s$ for all $a > b$, $q_s^{ab} = Q_s$ for all $a = b$ (same for the conjugate),
(iv) $b_s^a = b_s$ for all $a = 1, ..., r$.

Moreover, since we want to describe the minimizers 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:

$$Q - q = \delta q/\beta, \quad \hat{Q} - \hat{q} = -\beta\delta\hat{q}, \quad \hat{q} \sim \beta^2\hat{q}, \quad \hat{m} \sim \beta\hat{m}, \quad \hat{R} \sim \beta\hat{R}. \tag{A20}$$

b. Interaction term. We now proceed with the calculation of the different terms in Eq. (A16), where we can substitute the RS ansatz. In the interaction term, neglecting terms of $\mathcal{O}(n^2)$, we get

$$G_i = \exp\left(-n\left(\sum_s\left(\hat{m}_s m_s + \sum_c \hat{R}_{sc}R_{sc} + \frac{\hat{Q}_s Q_s}{2} - \frac{\hat{q}_s q_s}{2}\right)\right)\right). \tag{A21}$$

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

$$\log(G_i)/d = g_i = -\beta\left(\sum_s\left(\hat{m}_s m_s + \sum_c \hat{R}_{sc}R_{sc} + \frac{1}{2}(\hat{q}_s\delta q_s - \delta\hat{q}_s q_s)\right)\right). \tag{A22}$$

3. 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 (A18) to get

$$\begin{aligned} G_S = &\int \mathcal{D}T_+ \int \mathcal{D}T_- \exp\left(\sum_c \hat{Q}_T^c T_c^2 + \sum_c \hat{m}_T^c T_c + \hat{q}_T T_+ T_-\right) \int \prod_{s,a} d\mu\left(w_s^a\right) e^{-\frac{\gamma}{2}(w_1^a - w_2^a)^2} \\ &\times \prod_s \exp\left(\hat{m}_s \sum_a w_s^a + \frac{1}{2}(\hat{Q}_s - \hat{q}_s)\sum_a \left(w_s^a\right)^2 + \frac{1}{2}\hat{q}_s\left(\sum_a w_s^a\right)^2 + \sum_c \hat{R}_{sc}\sum_a w_s^a T_c\right). \end{aligned} \tag{A23}$$

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 $\hat{q}_T T_+ T_- = \frac{\hat{q}_T}{2}(T_+ + T_-)^2 - \frac{\hat{q}_T}{2}(T_+^2 + T_-^2)$, 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:

$$\begin{aligned} = &\int \mathcal{D}\tilde{z} \int \prod_s \mathcal{D}z_s \int \frac{dT_c}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_c \left(1 - 2\hat{Q}_T^c + \hat{q}_T\right)T_c^2 + \sum_c \left(\hat{m}_T^c + \sqrt{\hat{q}_T}\tilde{z}\right)T_c\right) \int \prod_a \mathcal{D}z_{12}^a \\ &\times \int \prod_{s,a} d\mu\left(w_s^a\right) \prod_s \exp\left(\frac{1}{2}(\hat{Q}_s - \hat{q}_s)\sum_a \left(w_s^a\right)^2 + \left(\hat{m}_s + \sum_c \hat{R}_{sc}T_c + \sqrt{\hat{q}_s}z_s + is\sqrt{\gamma}z_{12}^a\right)\sum_a w_s^a\right). \end{aligned} \tag{A24}$$

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

$$g_S = A + \int \prod_s \mathcal{D}z_s \int \prod_c \mathcal{D}T_c \int \mathcal{D}\tilde{z} \log \int \mathcal{D}z_{12} \int \prod_s d\mu(w_s) \exp\left(\frac{1}{2}(\hat{Q}_s - \hat{q}_s)w_s^2 + B_s w_s\right), \tag{A25}$$

where

$$A = \frac{\sum_c \left(\hat{m}_T^c\right)^2\left(1 - 2\hat{Q}_T^{-c}\right) + 2\hat{q}_T\left(\sum_c \hat{m}_T^c\right)^2}{2\left((1 - 2\hat{Q}_T^+)(1 - 2\hat{Q}_T^-) - \hat{q}_T^2\right)}, \tag{A26}$$

$$B_s = b_s(T_\pm, z_\pm, \tilde{z}, z_s) + is\sqrt{\gamma}z_{12}, \tag{A27}$$

$$b_s = \hat{m}_s + \sqrt{\hat{q}_s}z_s + \sum_c \left[m_T^c\hat{R}_{sc} + \frac{\hat{R}_{sc}}{\sqrt{\left(1 - 2\hat{Q}_T^c + \hat{q}_T\right)}}T_c + \frac{\sqrt{\hat{q}_T}}{\sqrt{1 - \sum_{c'}\frac{\hat{q}_T}{\left(1 - 2\hat{Q}_T^{c'} + \hat{q}_T\right)}}}\frac{\hat{R}_{sc}}{\left(1 - 2\hat{Q}_T^c + \hat{q}_T\right)}\tilde{z}\right]. \tag{A28}$$

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

$$g_S = A + \int \prod_s \mathcal{D}z_c \int \prod_c \mathcal{D}T_c \int \mathcal{D}\tilde{z} \log \int \mathcal{D}z_{12} \exp\left(\sum_s \max_{w_s}\left(-\frac{\lambda + \delta\hat{q}_s}{2} w_s^2 + B_s w_s\right)\right), \tag{A29}$$

and the maximization gives

$$w_s^\star = \frac{B_s}{(\lambda + \delta\hat{q}_s)}, \quad \max_{w_s}\left(-\frac{\lambda + \delta\hat{q}_s}{2} w_s^2 + B_s w_s\right) = \frac{B_s^2}{2(\lambda + \delta\hat{q}_s)}. \tag{A30}$$

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

(i) Expand the sums in Eq. (A29).

(ii) Perform the z_{12} Gaussian integration and take the log of the result.

(iii) 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.

(iv) Perform the remaining Gaussian integrations.

(v) Use identities (A6) and (A7) to remove the dependence on the conjugate fields appearing in the teacher measure and only retain a dependence on m_T^c, Q_T^c, and q_T.

The final expression reads

$$\begin{aligned} g_S = \frac{\beta}{2\left(\prod_s(\lambda + \gamma + \delta\hat{q}_s) - \gamma^2\right)} \Bigg[& \left(\sum_s \left(\hat{m}_s + \sum_s m_T^c \hat{R}_{sc}\right)^2 (\lambda + \gamma + \delta\hat{q}_{\neg s}) + 2\gamma \prod_s \left(\hat{m}_s + \sum_c m_T^c \hat{R}_{sc}\right)\right) \\ & + \left(\sum_s \hat{q}_s(\lambda + \gamma + \delta\hat{q}_{\neg s})\right) + \left(\sum_c \left(1 - \left(m_T^c\right)^2\right)\left(\sum_s \hat{R}_{sc}^2(\lambda + \gamma + \delta\hat{q}_{\neg s}) + 2\gamma \prod_s \hat{R}_{sc}\right)\right) \\ & + \left(2\left(q_T - m_T^+ m_T^-\right)\left(\sum_s \left(\prod_c \hat{R}_{sc}(\lambda + \delta\hat{q}_{\neg s})\right) + \gamma\left(\prod_c \left(\sum_s \hat{R}_{sc}\right)\right)\right)\right)\Bigg], \end{aligned} \tag{A31}$$

where the notation $\neg s$ denotes the other student index with respect to the one used in the corresponding sum or product.

c. 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 subsection 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. (A19) we get

$$\begin{aligned} G_E = & \int \frac{du d\hat{u}}{2\pi} e^{iu\hat{u}} \int \prod_a \left(\frac{d\lambda^a d\hat{\lambda}^a}{2\pi} e^{i\lambda^a \hat{\lambda}^a}\right) e^{-\frac{\Delta}{2}\sum_{ab}\hat{\lambda}_a\hat{\lambda}_b q - \Delta\hat{u}R\sum_a \hat{\lambda}_a - \frac{\Delta}{2}(\hat{u})^2 q_T} \\ & \times \prod_a e^{-\beta\,\ell(u + c\tilde{m} + \tilde{b}, \lambda^a + cm + b)}. \end{aligned} \tag{A32}$$

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 factorizes over the replica index. Up to first order in r when $r \to 0$, we find for $g_E = \log G_E/d$:

$$g_E = \int \mathcal{D}z \int \mathcal{D}u \log \int \mathcal{D}\lambda e^{-\beta\,\ell(\sqrt{\Delta q_T}u + c\tilde{m} + \tilde{b}, \sqrt{\Delta(Q-q)}\lambda + \frac{\sqrt{\Delta}R}{\sqrt{q_T}}u + \sqrt{\Delta\frac{(q-R^2)}{q_T}}z + cm + b)}, \tag{A33}$$

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

$$\log \int \mathcal{D}\lambda e^{-\beta\,\ell(\sqrt{\Delta q_T}u + c\tilde{m} + \tilde{b}, \sqrt{\Delta(Q-q)}\lambda + \frac{\sqrt{\Delta}R}{\sqrt{q_T}}u + \sqrt{\Delta\frac{(q-R^2)}{q_T}}z + cm + b)} = -\beta M, \tag{A34}$$

with

$$M = \min_\lambda \frac{\lambda^2}{2} + \ell\left(\sqrt{\Delta q_T}u + c\tilde{m} + \tilde{b}, \sqrt{\Delta\delta q}\lambda + \frac{\sqrt{\Delta}R}{\sqrt{q_T}}u + \sqrt{\Delta\frac{(q-R^2)}{q_T}}z + cm + b\right). \tag{A35}$$

To simplify further, we can shift $\frac{\sqrt{\Delta}R}{\sqrt{q_T}}u + \sqrt{\Delta\frac{(q-R^2)}{q_T}}z \to \sqrt{\Delta q}z'$. Then, given the definition of the logistic loss (A2), we can split the u integration over the intervals $\sqrt{\Delta q_T}u + c\tilde{m}_c > 0$ and $\sqrt{\Delta q_T}u + c\tilde{m}_c < 0$ and eventually get (reestablishing the s, c indices):

$$g_E(s,c) = \sum_y \int \mathcal{D}z H\left(-y\frac{q_s\frac{cm_T^c+b_T^c}{\sqrt{1}} + \sqrt{\Delta_c}R_{sc}z}{\sqrt{\Delta_c(q_s - R_{sc}^2)}}\right) M_E(y,s,c), \tag{A36}$$

where $H(x) = \frac{1}{2}\,\mathrm{erfc}(x/\sqrt{2})$ is the Gaussian tail function and we defined the proximal as

$$M_E(y,s,c) = \max_\lambda -\frac{\lambda^2}{2} - \ell(y, \sqrt{\Delta_c\delta q_s}\lambda + \sqrt{\Delta_c q_s}z + cm_s + b_s). \tag{A37}$$

Note that this simple 1D optimization 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 $\mathcal{W}_{c,y}$ 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

$$\mathcal{L}(\mathcal{D}) = \sum_{c=\pm}\sum_{y=0,1}(\mathcal{W})_{(c,y)}Y_y \log P(\hat{Y}_y). \tag{A38}$$

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

$$\mathcal{W} = 2\begin{pmatrix} w_+w_1 & w_+(1-w_1) \\ (1-w_+)w_1 & (1-w_+)(1-w_1)\end{pmatrix}, \tag{A39}$$

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 optimization of the proximal would be affected.

4. 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 Eq. (A16) is dominated by the typical configuration of the order parameters, which is found by extremizing the free entropy with respect to all the overlap parameters:

$$\Phi_W = \underset{o.p.}{\mathrm{extr}}\left\{g_I + g_S + \sum_{s,c}\alpha_{s,c}\,g_E(s,c)\right\}. \tag{A40}$$

The saddle point is typically found by fixed-point 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, $m_T^\pm$ and q_T, and the pattern densities $\alpha_{s,c}$. In the main text, for simplicity, we parametrize $\alpha_{s,c}$ 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 $\mathcal{D}_1$.

d. 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, then probability of obtaining an output $\hat{y}$ from the trained model s is given by

$$P(Y=y,\hat{Y}=\hat{y}) = \mathbb{E}_{\mathbf{x}(c)}\left\langle\Theta\left(y\left(\frac{\mathbf{W}_T^c\cdot\mathbf{x}(c)}{\sqrt{d}} + \tilde{b}\right)\right)\Theta\left(\hat{y}\left(\frac{\mathbf{W}_s\cdot\mathbf{x}(c)}{\sqrt{d}} + b\right)\right)\right\rangle_{\mu(\mathbf{W}_T,\mathbf{W})} \tag{A41}$$

$$= \mathbb{E}_{\mathbf{x}(c)}\left\langle\int\frac{du\,d\hat{u}}{2\pi}e^{i\hat{u}(u-\sum_{i=1}^d\frac{W_{T,i}x_i}{\sqrt{d}})}\int\frac{d\lambda\,d\hat{\lambda}}{2\pi}e^{i\hat{\lambda}(\lambda-\sum_{i=1}^d\frac{W_ix_i}{\sqrt{d}})}\right\rangle\Theta(y(u+\tilde{b}))\Theta(\hat{y}(\lambda+b)), \tag{A42}$$

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:

$$\mathbb{E}_{\boldsymbol{x}(c)}\left\langle e^{-i\hat{\lambda}\frac{\boldsymbol{W}_s\cdot\boldsymbol{x}(c)}{\sqrt{d}}-i\hat{u}\frac{\boldsymbol{W}_T\cdot\boldsymbol{x}(c)}{\sqrt{d}}}\right\rangle = e^{-ic(\hat{\lambda}m+\hat{u}\tilde{m})}e^{-\frac{\Delta}{2}(\hat{\lambda}^2Q+2\hat{u}\hat{\lambda}R+\hat{u}^2)}. \tag{A43}$$

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 in the $\beta \to \infty$ limit they are equal up to the first order.

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

$$P(Y = y, \hat{Y} = \hat{y}) = \int_{-\infty}^{\infty} \mathcal{D}u\Theta\big(y\big(\sqrt{\Delta_c}u + cm_T^c + b_T^c\big)\big)H\left(-\hat{y}\frac{\sqrt{\Delta_c}R_{sc}u + cm_s + b_s}{\sqrt{\Delta_c(q_s - R_{sc}^2)}}\right). \tag{A44}$$

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

$$P(Y = 1) = \rho H\left(-\frac{m_T^+ + b_T^+}{\sqrt{\Delta_+}}\right) + (1-\rho)H\left(\frac{m_T^- - b_T^-}{\sqrt{\Delta_-}}\right), \tag{A45}$$

and the generalization error

$$\epsilon_g = \int_{-\infty}^{\infty} \mathcal{D}uH\left(\mathrm{sign}\big(\big(\sqrt{\Delta_c}u + cm_T^c + b_T^c\big)\big)\frac{\sqrt{\Delta_c}R_{sc}u + cm_s + b_s}{\sqrt{\Delta_c\big(q_s - R_{sc}^2\big)}}\right). \tag{A46}$$

5. Parameters used in the figures

The following list contains the parameters of the T-M model used to plot the figures of the paper.

(i) Fig. 1(d): $\Delta_+ = 0.5$, $\Delta_- = 2_0.5$, $\alpha = 2.5$, $q_T - 0.2$.
(ii) Fig. 3(a): $m_\pm = 0.2$, $\alpha = 0.5$, $\Delta_+ = 0.5$, $\Delta_- = 0.5$, $b_+ = 0$, $b_- = 0$.
(iii) Fig. 3(b): $\alpha = 0.5$, $\Delta_+ = 0.5$, $\Delta_- = 0.5$, $b_+ = 0$, $b_- = 0$.
(iv) Fig. 4: $\alpha = 0.5$, $q_T = 1$, $m = 0.5$, $b_+ = 0$, $b_- = 0$.
(v) Fig. 5: $\rho = 0.1$, $m = 0.2$, $\Delta_+ = 0.5$, $\Delta_- = 0.5$, $b_+ = 0$, $b_- = 0$.
(vi) Fig. 6(a): $\rho = 0.1$, $q_T = 0.8$, $\Delta_+ = 2.0$, $\Delta_- = 0.5$, $\alpha = 0.5$, $m_+ = 0.3$, $m_- = 0.1$, $b_+ = 0.5$, $b_- = 0.5$.

APPENDIX B: REAL DATA VALIDATION

In the next two sections, we demonstrate the ability of the teacher-mixture model to mimic unfairness scenarios in real-world applications. In particular, we perform this validation through a set of numerical experiments on the CelebA dataset [19]. This dataset consists of a collection of face images of celebrities, equipped with metadata indicating the presence of specific attributes in each picture. As can be seen in Fig. 9, the consistent amount of these attributes allows one to explore many possible learning scenarios in unfairness conditions. This feature of CelebA together with its size and the high-dimensional nature of face pictures, makes it a good candidate for validating the teacher-mixture model on real datasets. Moreover, as shown in Fig. 8 through a PCA clustering, the different subpopulations associated to a given CelebA attribute are overlapping and hard to disentangle. This situation precisely corresponds to the high-noise regime the teacher-mixture model is meant to describe. Interestingly, the picture emerging from the simulations on CelebA turned out to be quite general and further extendable to lower-dimensional datasets such as the Medical Expenditure Panel Survey (MEPS) dataset [58]. More details on both datasets are discussed in Appendixes B 2 and C 3. Here we provide a general overview on the experimental framework applied to CelebA.

1. Model motivation

We construct a dataset by subsampling CelebA and by preprocessing the selected images through an Xception network [59] trained on ImageNet [60]. As depicted in the scatter plot in Fig. 7, 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 subpopulations. For example, in the figure we color 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" [61]. 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. 7 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 generalization performance of the model when evaluated on the different subpopulations. The presence of a gap is a clear indicator of unfairness, induced by an implicit bias toward 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 and unbiased classifier is a very

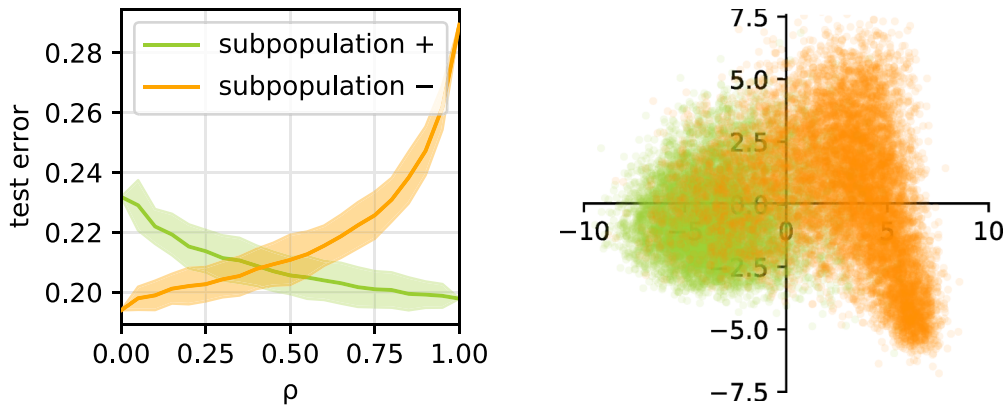


FIG. 7. Relative representation and bias. Numerical experiments on a subsample of the CelebA dataset. (Left) A 2D projection of the preprocessed dataset, obtained from PCA, where the colors represent the two subpopulations. (Right) Per community test error, as the fraction of samples from the two subpopulations is varied (dataset dimension is fixed).

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 data points and the balance between the subpopulations 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 recognized by looking at the joint distribution of inputs and labels. For example, the balance between the positive and 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 labeled differently. The present work aims at modeling the data structure observed in these types of experiments, to obtain detailed understanding of the various sources of bias in these problems.

2. 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) [19]. To obtain the results presented in the main text we apply the following preprocessing 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 [62], we then

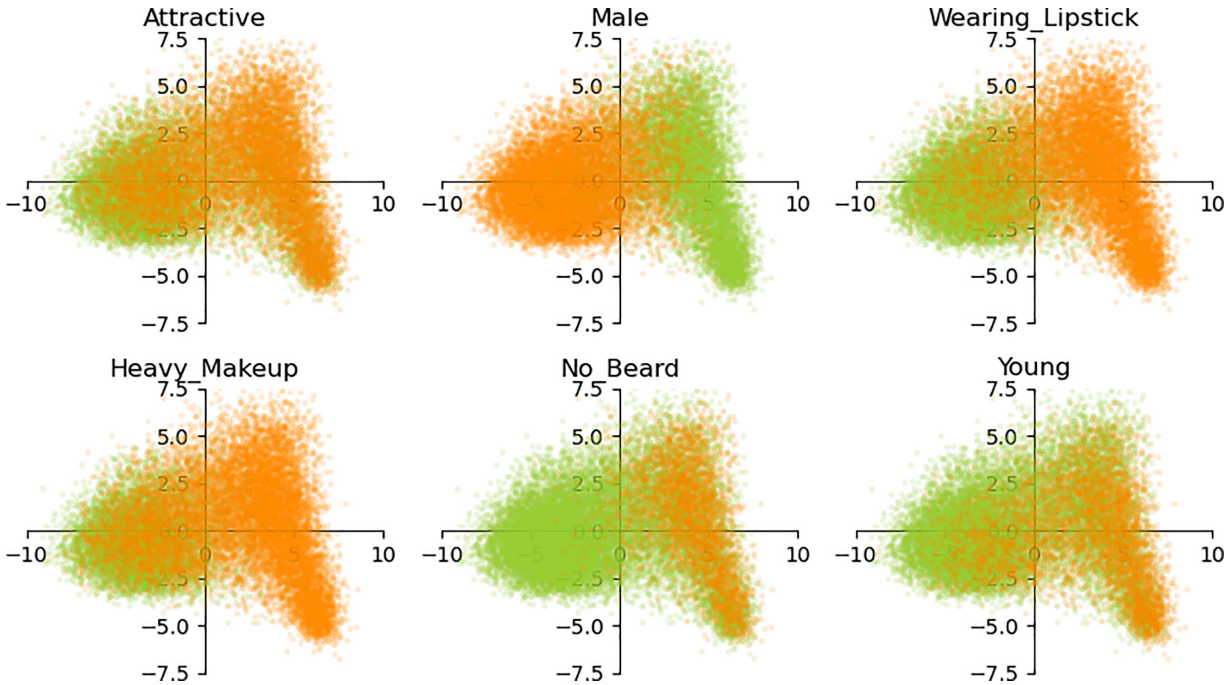


FIG. 8. Clustering CelebA according to attributes. We show 6 of the 40 attributes in CelebA demonstrating a neat clustering.

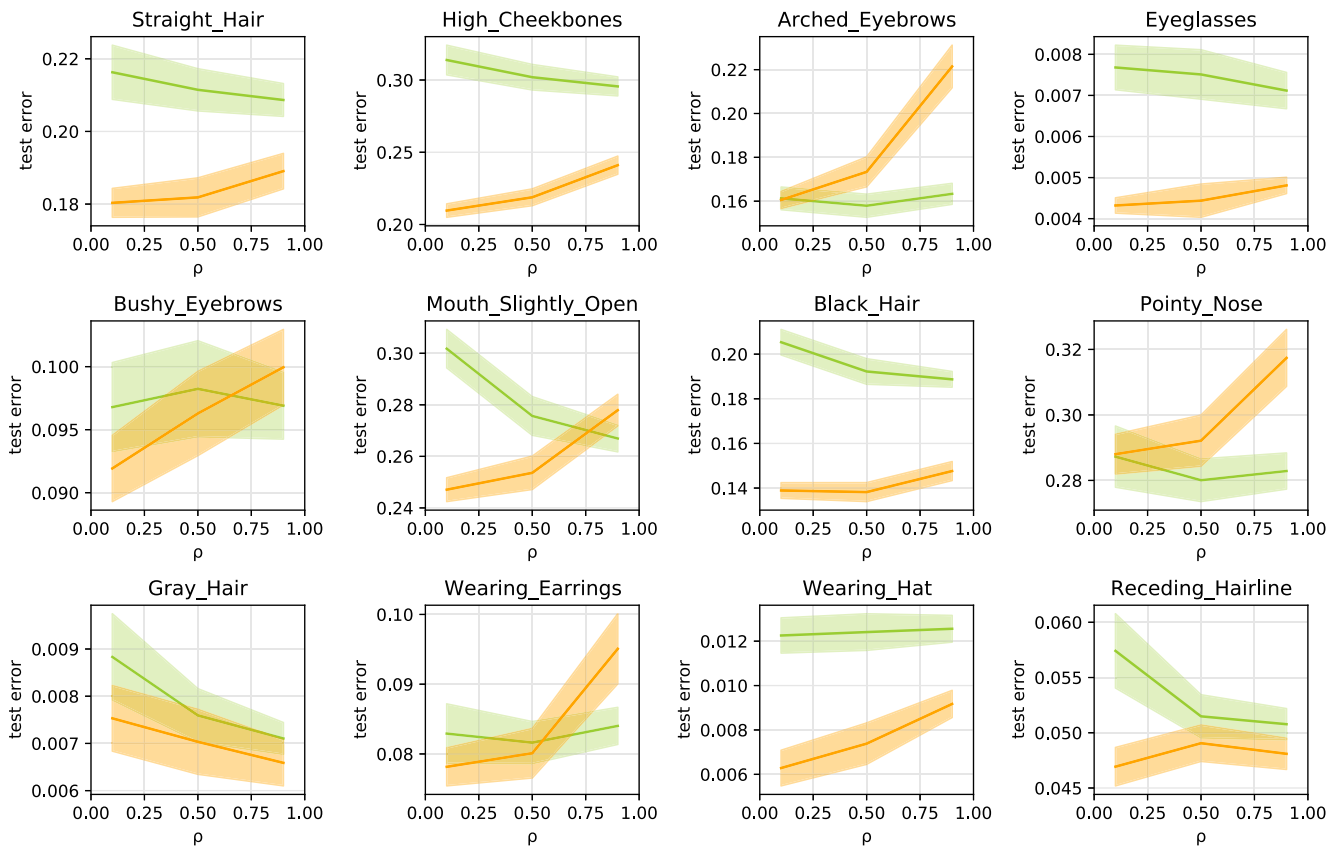


FIG. 9. Relative representation across attributes. The panels show the generalization error depending on the relative representation in different attributes. The subpopulations + (green) − (orange) are obtained splitting according to the attribute "Wearing_Lipstick." The simulations are averaged over 100 samples.

preprocess the dataset using the features extracted from an Xception convolutional network [59] pretrained on Imagenet [60]. Finally, we collect the extracted features together with the associated binary attributes in a json file.

By applying PCA on the preprocessed 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. 8, where we use colors 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. 8, the two subpopulations are overlapping

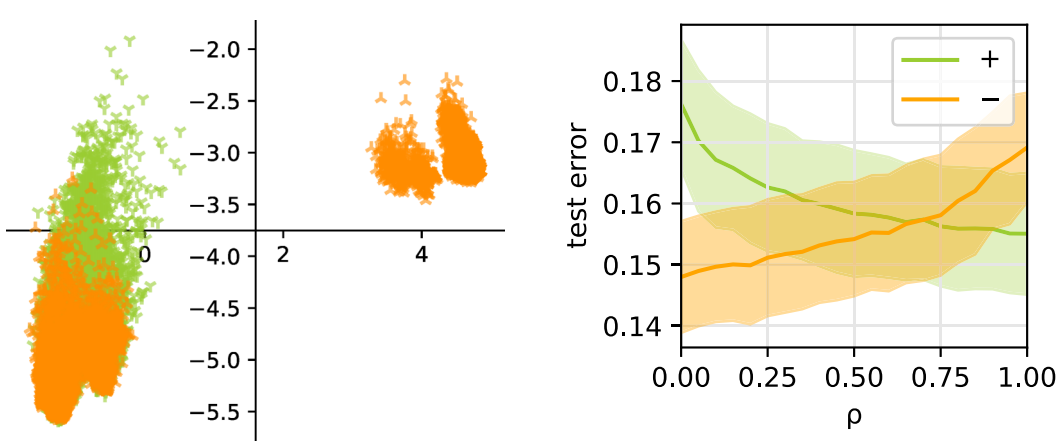


FIG. 10. MEPS dataset. (Left) Clustering in the MEPS dataset, according to be above or below the average age. (Right) Crossing of the generalization error as the relative representation ρ is changed. The simulations are averaged over 100 samples.

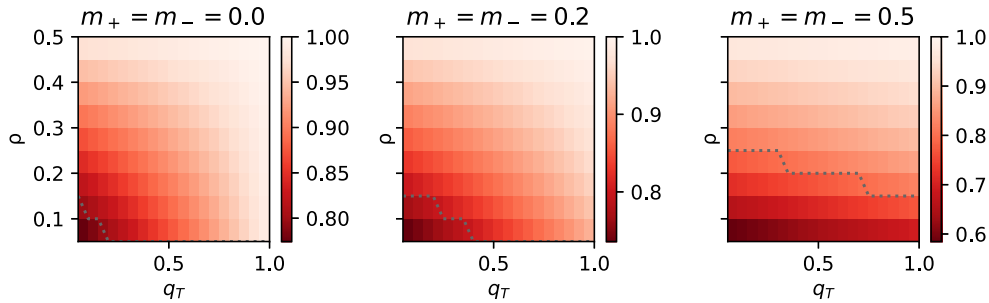


FIG. 11. Bias with two different rules to be learned. The three phase diagrams give the DI depending on ρ (y axis) and q_T (x axis). Moving from the left panel to the right panel m_+ and m_i are increased. The other parameters are: $\alpha = 0.5$, $\Delta_+ = 0.5$, $\Delta_- = 0.5$, $b_+ = 0$, $b_- = 0$.

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. 8, 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.

Anyone of the other attributes can be considered as a possible target, and thus be used to label the data points. The final preprocessing step consists in downsampling further the data 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 Appendix C we comment more on the bias induced by different label ratios.

As Fig. 9 illustrates, there is a large number of possible outcomes concerning the behavior 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 behaviors 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. 9, we train a classifier with logistic regression and L_2-regularization. In particular, we use the LogisticRegression class from scikit-learn [63]. 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 hyperparameter 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.

3. Other datasets

The observations made on the CelebA dataset are quite general and can be further extended to lower-dimensional datasets. As example of this, we considered the MEPS dataset. This is a dataset containing a large set of surveys which have been conducted across the United States 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. 10, the behavior 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 subpopulations, according to the middle point of the age distribution. Moreover, the generalization error per community exhibits a crossing according to the relative representation.

APPENDIX C: EXPLORATION OF THE PARAMETER SPACE

Supporting results

This section presents supporting results on the sources of bias. In Fig. 11, we repropose 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 $m_+ = m_- = 0$, then when the rules become identical ($q_T = 1$) the bias is removed. However, if $m_+ = m_- \neq 0$, then 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.

The main difference with respect to the case with $q_T \neq 1$ is that, if $q_T = 1$, then 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 subpopulation, and this effect can be cured by increasing in α. This is illustrated in Fig. 12, which 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.

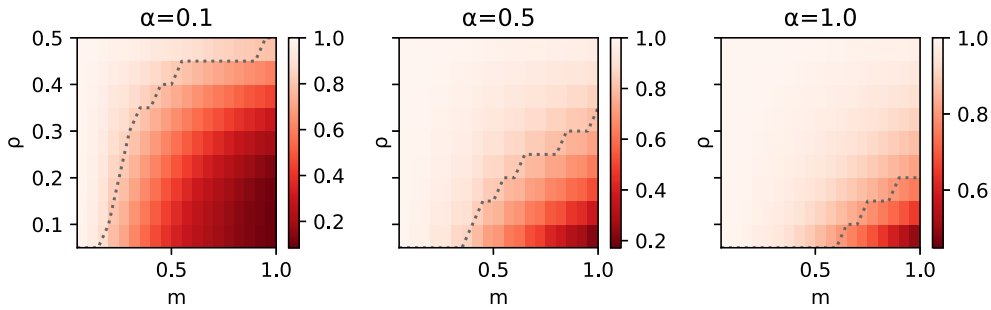


FIG. 12. Bias with a learnable rule. We show the accuracy gain as function of the proportion of group + (ρ) and the correlation between label and group (m_+, m_-). The different figures show how of increasing the dataset size (increasing from left to right) mitigates the bias. The other parameters are: $q_T = 1.0$, $\Delta_+ = 0.5$, $\Delta_- = 0.5$, $b_+ = 0$, $b_- = 0$.

The results shown until this point are agnostic with respect to the relative fraction of labels inside the subpopulations. 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 toward 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. 13 to exemplify this concept.

When the subpopulations are equally represented $\rho = 0.5$, the separations between bias toward + 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

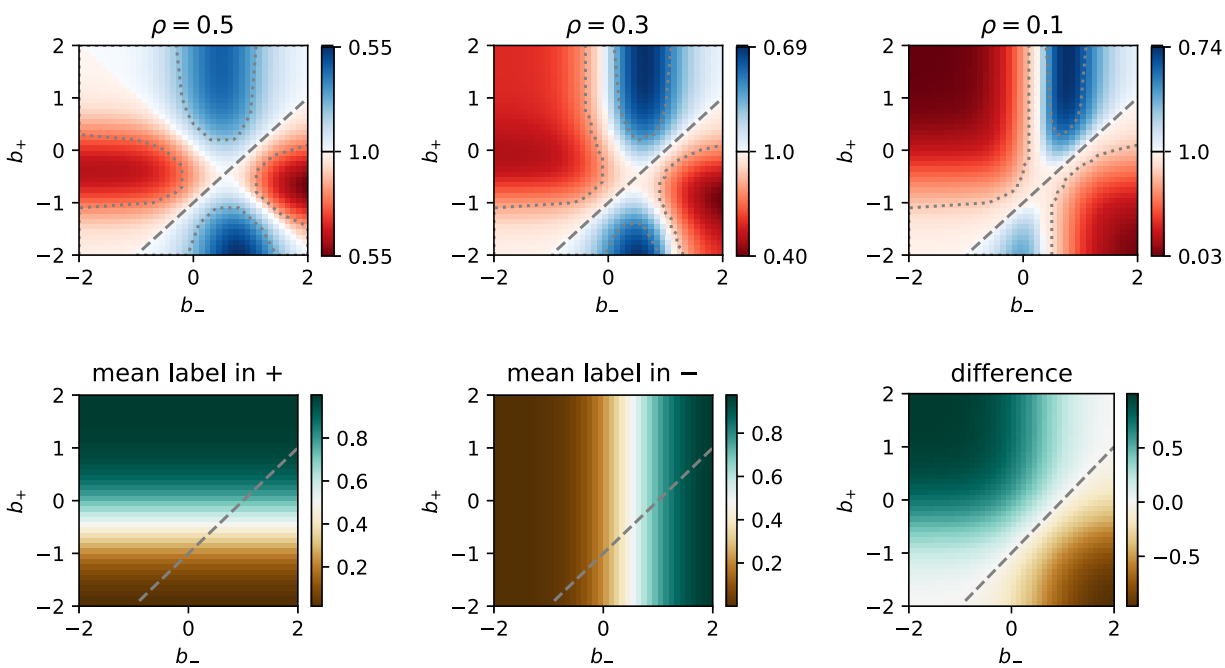


FIG. 13. Labels within groups and classifier bias. The *first row* shows the DI as faction of b_+ and b_- with $\Delta_+ = \Delta_- = 0.5$, $\alpha = 0.5$, $m_+ = m_- = 0.5$. From left to right, the relative representation ρ moves from equally represented groups to having group + under-represented. The 80% threshold is denoted by the dotted line. The dashed line indicates equal within-group label fraction. The *second row* shows the average labeling in + (left), − (center), and their difference (right). Notice that these diagrams are independent of ρ and therefor apply to the three settings shown in the first row.

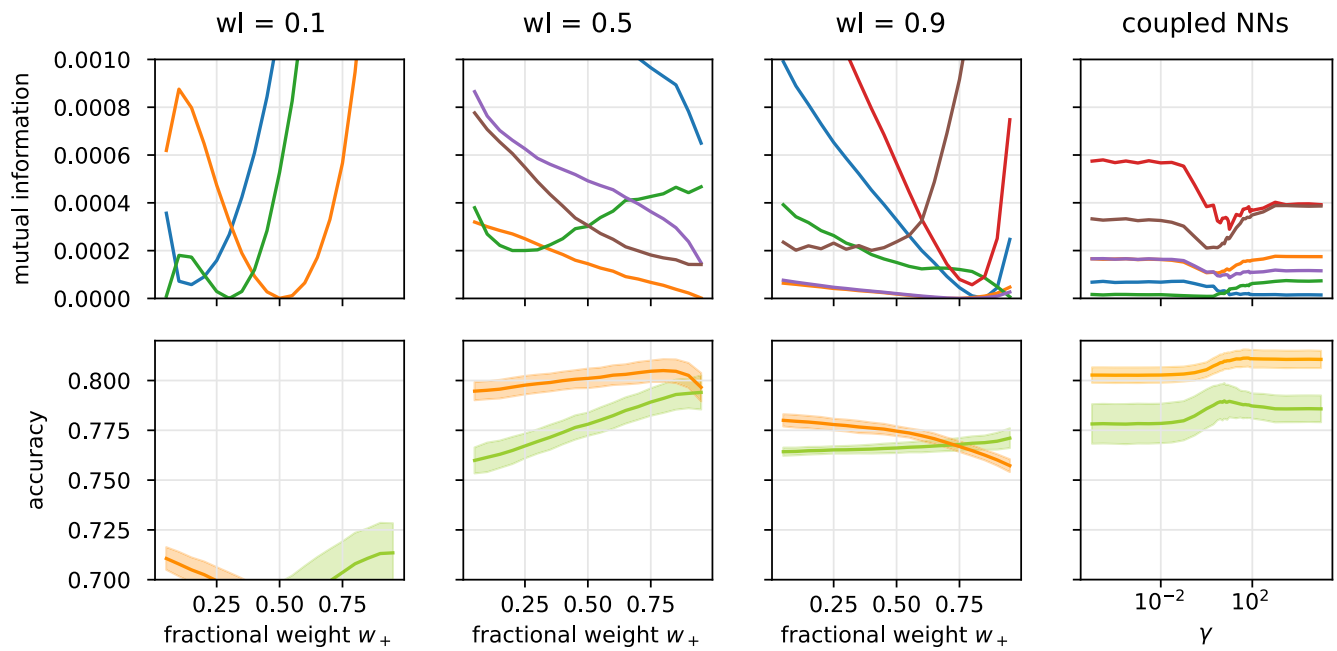


FIG. 14. Mitigation using re-weighting on real data. The four panels show the same quantities as in Fig. 6(a) but applied to the CelebA dataset. Each panels shows in upper figure the mutual information on the fairness metrics—statistical parity, equal opportunities, equal accuracy, equal odds, predicted parity 1, and predicted parity 10—and in the lower figure the accuracy of for subpopulation + and subpopulation −. The first three panels show the effect on the reweighing strategy for different values of the label weight (from left to right $w_l = 0.1, 0.5, 0.9$). The last panel shows the performance of the coupled neural network strategy.

the separation line is distorted. Finally, observe that the line of equal label fraction (bottom right panel) is not centered in the diagram because $m_+ = m_- \neq 0$.

APPENDIX D: MITIGATION STRATEGIES

1. Real data

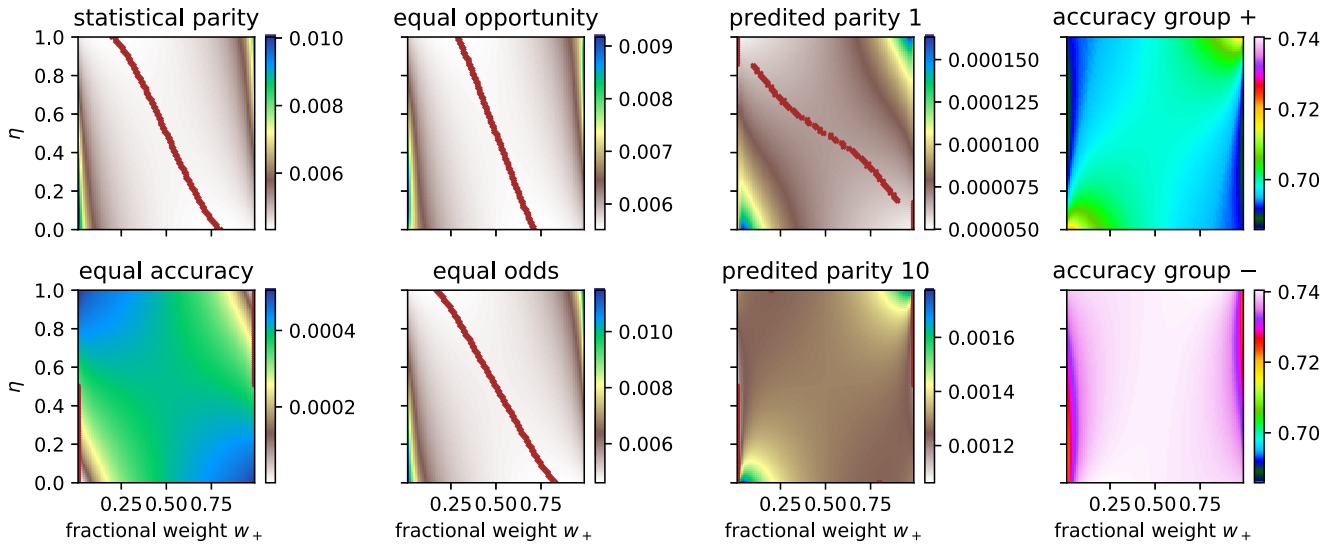
In Fig. 6 of the main text, we show the effect of reweighing in the synthetic model. The same analysis can be applied to real data, yielding similar results. In particular, in line with the other validations, we present in Fig. 14 the result for the CelebA dataset when the splitting is done according to the "Wearing_Lipstick" and the target feature is "Wavy_Hair."

Similarly to what observed in the synthetic dataset, the coupled neural network strategy allows for a better performance on all the fairness metrics while retraining a high accuracy for both subpopulations.

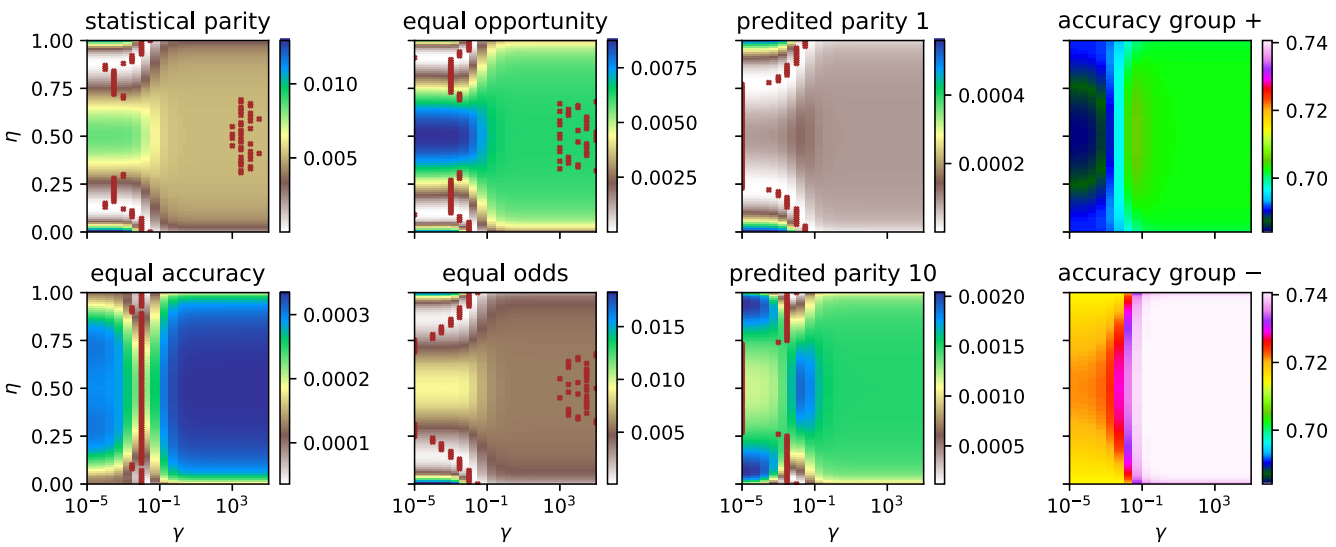
2. Additional results varying group membership

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. 15(a) 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 subpopulation is greatly deteriorated.

We can understand the larger picture by looking at the different fairness metrics described in the main text, Fig. 15(b), for which the same observations apply. Since η is not an actual hyperparameter, 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 optimized all at once.



(a) **MI with errors in the group membership under community re-weighting strategy.**



(b) **MI with errors in the group membership under coupled neural networks strategy.**

FIG. 15. Effect of noise in the attribute of the subcommunities. In the heatmaps we show in colors how having imperfect information concerning the subcommunity membership affect each fairness metrics (six left figures) and the accuracy (right two plots). The vertical axis of the figures represents the probability of mismatch η, while the horizontal axis refer to the parameter of the strategy (w_+ and γ, respectively). For every value of the mismatch probability, we denote with red points the minima of the mutual information for each fairness metrics.

[1] J. Buolamwini and T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (PMLR, 2018), pp. 77–91.

[2] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh *et al.*, Ethical and social risks of harm from language models, arXiv:2112.04359.

[3] R. Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Polity, Cambridge, UK, 2019).

[4] S. U. Noble, *Algorithms of Oppression* (New York University Press, New York, NY, 2018).

[5] V. Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, New York, NY, 2018).

[6] M. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, Cambridge, MA, 2018).

[7] H. Suresh and J. V. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in *Proceedings of the ACM Conference on Equity and Access*

in Algorithms, Mechanisms, and Optimization (EAAMO'21), Virtual Event, October 5–9, 17:1 (ACM, New York, NY, 2021).

[8] C. C. Perez, *Invisible Women: Data Bias in a World Designed for Men* (Abrams, New York, NY, 2019).

[9] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, Handling imbalanced datasets A review, GESTS Int. Trans. Comput. Sci. Eng. **30**, 25 (2006).

[10] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, Review of classification methods on unbalanced data sets, IEEE Access **9**, 64606 (2021).

[11] X.-w. Chen and M. Wasikowski, FAST: A ROC-based feature selection metric for small samples and imbalanced data classification problems, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2008), pp. 124–132.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. **16**, 321 (2002).

[13] X.-Y. Liu, J. Wu, and Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. Part B Cybern. **39**, 539 (2008).

[14] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, Towards fairness in visual recognition: Effective strategies for bias mitigation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2020), pp. 8919–8928.

[15] B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz, Simple data balancing achieves competitive worst-group-accuracy, in *Proceedings of the Conference on Causal Learning and Reasoning* (PMLR, 2022), pp. 336–351.

[16] R. N. Hughes, Sex does matter: Comments on the prevalence of male-only investigations of drug effects on rodent behaviour, Behav. Pharmacol. **18**, 583 (2007).

[17] R. N. Hughes, Sex still matters: Has the prevalence of male-only studies of drug effects on rodent behaviour changed during the past decade? Behav. Pharmacol. **30**, 95 (2019).

[18] L. Zdeborová and F. Krzakala, Statistical physics of inference: Thresholds and algorithms, Adv. Phys. **65**, 453 (2016).

[19] Z. Liu, P. Luo, X. Wang, and X. Tang, Deep learning face attributes in the wild, in *Proceedings of International Conference on Computer Vision (ICCV)* (IEEE Computer Society, Piscataway, NJ, USA, 2015).

[20] F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova, The role of regularization in classification of high-dimensional noisy gaussian mixture, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 6874–6883.

[21] L. Saglietti and L. Zdeborová, Solvable model for inheriting the regularization through knowledge distillation, in *Proceedings of the Conference on Mathematical and Scientific Machine Learning* (PMLR, 2022), pp. 809–846.

[22] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, Certifying and removing disparate impact, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery (ACM), New York, NY, USA, 2015), pp. 259–268.

[23] U. E. E. O. Commission *et al.*, *Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection procedures* (U.S. Equal Employment Opportunity Commission, Washington, DC, 1979).

[24] P. Charbonneau, E. Marinari, M. Mézard, G. Parisi, F. Ricci-Tersenghi, G. Sicuro, and F. Zamponi, *Spin Glass Theory and Far Beyond* (World Scientific, Singapore, 2023).

[25] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, Singapore, 1987), Vol. 9.

[26] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).

[27] C. Thrampoulidis, S. Oymak, and B. Hassibi, Regularized linear regression: A precise analysis of the estimation error, in *Proceedings of the Conference on Learning Theory* (PMLR, 2015), pp. 1683–1709.

[28] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová, Learning curves of generic features maps for realistic datasets with a teacher-student model, Adv. Neural Info. Process. Syst. **34**, 18137 (2021).

[29] F. Gerace, L. Saglietti, S. S. Mannelli, A. Saxe, and L. Zdeborová, Probing transfer learning with a model of synthetic correlated datasets, Mach. Learn.: Sci. Technol. **3**, 015030 (2022).

[30] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. 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 and Data Mining* (Association for Computing Machinery (ACM), New York, NY, USA, 2018), pp. 2239–2248.

[31] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, A clarification of the nuances in the fairness metrics landscape, Sci. Rep. **12**, 4209 (2022).

[32] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, A reductions approach to fair classification, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2018), pp. 60–69.

[33] A. Agarwal, M. Dudík, and Z. S. Wu, Fair regression: Quantitative definitions and reduction-based algorithms, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2019), pp. 120–129.

[34] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, Classification with fairness constraints: A meta-algorithm with provable guarantees, in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), pp. 319–328.

[35] J. Kleinberg, S. Mullainathan, and M. Raghavan, Inherent trade-offs in the fair determination of risk scores, in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference, ITCS 2017*, Berkeley, CA, USA (Schloss Dagstuhl Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2017), pp. 3315–3323.

[36] S. Corbett-Davies and S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, J. Mach. Learn. Res. **24**, 3121 (2023).

[37] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning* (2019).

[38] F. Kamiran and T. Calders, Data preprocessing techniques for classification without discrimination, Knowl. Inf. Syst **33**, 1 (2012).

[39] D. Plecko and N. Meinshausen, Fair data adaptation with quantile preservation, J. Mach. Learn. Res. **21**, 1 (2020).

[40] K. Lum and J. Johndrow, A statistical framework for fair predictive algorithms, arXiv:1610.08077.

[41] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, Progressive neural networks, arXiv:1606.04671.

[42] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, Fairness through awareness, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Association for Computing Machinery (ACM), New York, NY, USA, 2012), pp. 214–226.

[43] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, Algorithmic decision making and the cost of fairness, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery (ACM), New York, NY, USA, 2017), pp. 797–806.

[44] M. Hardt, E. Price, and N. Srebro, Equality of opportunity in supervised learning, Adv. Neural Info. Process. Syst. **29**, 3315 (2016).

[45] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, Fairness constraints: Mechanisms for fair classification, in *Proceedings of the Conference on Artificial Intelligence and Statistics* (PMLR, 2017), pp. 962–970.

[46] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data **5**, 153 (2017).

[47] F. Zenke, B. Poole, and S. Ganguli, Continual learning through synaptic intelligence, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2017), pp. 3987–3995.

[48] L. Saglietti, S. S. Mannelli, and A. Saxe, An analytical theory of curriculum learning in teacher-student networks, J. Stat. Mech: Theory Exp. (2022) 114014.

[49] T. Calders and S. Verwer, Three naive Bayes approaches for discrimination-free classification, Data Min. Knowl. Discovery **21**, 277 (2010).

[50] R. H. Myers and R. H. Myers, *Classical and Modern Regression with Applications* (Duxbury Press, Belmont, CA, 1990), Vol. 2.

[51] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 2803–2813.

[52] The illustrated checkpoints 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.

[53] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, Phys. Rev. X **10**, 041044 (2020).

[54] U. Adomaityte, G. Sicuro, and P. Vivo, Classification of heavy-tailed features in high dimensions a superstatistical approach, Adv. Neural Info. Process. Syst. **36**, 43880 (2024).

[55] E. Cornacchia, F. Mignacco, R. Veiga, C. Gerbelot, B. Loureiro, and L. Zdeborová, Learning curves for the multi-class teacher–student perceptron, Mach. Learn.: Sci. Technol. **4**, 015019 (2023).

[56] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, An investigation of why overparameterization exacerbates spurious correlations, in *Proceedings of the International Conference on Machine Learning* (PMLR, 2020), pp. 8346–8356.

[57] S. J. Bell and L. Sagun, Simplicity bias leads to amplified performance disparities, in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery (ACM), New York, NY, USA, 2023), pp. 355–369.

[58] L. A. Blewett, J. A. Rivera Drew, R. Griffin, N. Del Ponte, and P. Convey, *IPUMS Health Surveys: Medical Expenditure Panel Survey, version 2.1* (IPUMS, Minneapolis, MN, 2021).

[59] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Piscataway, NJ, USA, 2017), pp. 1251–1258.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2009), pp. 248–255.

[61] To be mindful on the Ethical Considerations of using the CelebA datast, we do not use protected attributes like binary genders and age.

[62] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg *et al.*, TensorFlow: Large-scale Machine Learning on Heterogeneous Systems (tensorflow.org, 2015).

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel *et al.*, Scikit-learn machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).