Fairness for the People, by the People: Minority Collective Action

Omri Ben-Dov Samira Samadi Amartya Sanyal Alexandru Ţifrea

Abstract

Machine learning models often preserve biases present in training data, leading to unfair treatment of certain minority groups. Despite an array of existing firm-side bias mitigation techniques, they typically incur utility costs and require organizational buy-in. Recognizing that many such models rely on user-contributed data, we introduce a framework for minority Algorithmic Collective Action, where a coordinated minority group strategically relabels its own data to enhance fairness, without altering the firm's training process. We propose three practical, model-agnostic strategies to approximate ideal relabeling and validate them on five real-world datasets. Our findings show that with limited participation, a collective can substantially reduce unfairness with a small impact on overall accuracy.

1 Introduction

As machine learning (ML) tools become increasingly accessible, more firms deploy them for decision-making. However, ML models often perpetuate societal biases present in their training data, leading to unfair outcomes across demographic groups [1]. Moreover, most fairness-preserving learning algorithms incur a non-negligible cost in accuracy or computational resources [2, 3, 4, 5], which can discourage practical adoption.

For a running example, consider a recommendation system that predicts whether a user will engage with an item and then tailors suggestions accordingly. Since different demographic groups often exhibit distinct behavior patterns, the model implicitly learns correlations from group membership to engagement rates. As a result, underrepresented groups may receive systematically less relevant recommendations that reflect group behavior but fail to align with individual preferences.

Because firms control the training pipeline, end users lack access to these algorithms and cannot directly enforce fair treatment. Yet, affected users routinely generate and share data — through clicks, ratings, or other contributions — that is used to train the firm's models. Consequently, if underrepresented minority groups collaboratively alter the data they share, they might be able to steer the learned model towards fairer behavior, even without access to the firm's training pipeline.

This idea is reminiscent of *pre-processing* fairness techniques [6, 7, 8, 9], which modify the data before model training. Unlike these prior approaches, which assume centralized control over the data, we consider the setting of *algorithmic collective action* [10, 11, 12, 13, 14], in which a small group of users strategically modifies their own data to influence the correlations learned by the model.

We adapt the *erasure strategy* of Hardt et al. [10] to reduce predictive correlation between group membership and the target label by relabeling minority samples. The collective is restricted to members of the minority group since majority-group users may be less inclined to disrupt the status quo. We show that when a classifier is trained on data affected by this form of collective action, standard fairness metrics (e.g., demographic parity, equalized odds) improve substantially. This improvement is illustrated in Figure 1, where a small collective of minority users significantly reduces unfairness with minimal impact on prediction error.

The key obstacle in implementing the erasure strategy is that it requires knowledge of each user's label under a counterfactual group membership. Computing such counterfactual labels exactly would require access to an underlying causal model, which is typically infeasible in practice. To overcome this challenge, we propose three *model-agnostic* methods to estimate the counterfactual labels.

Our main contributions are: **(1)** We divert from the common firm-side fairness method and focus on user side by introducing the setting of **minority-only algorithmic collective action for fairness** in ML (Section 2), and design three algorithms that Pareto-dominate the random-choice baseline (Section 3). **(2)** Through experiments on benchmark datasets, we demonstrate that these algorithms can **significantly improve fairness metrics** with only a slight accuracy penalty and few label flips, and minimal knowledge

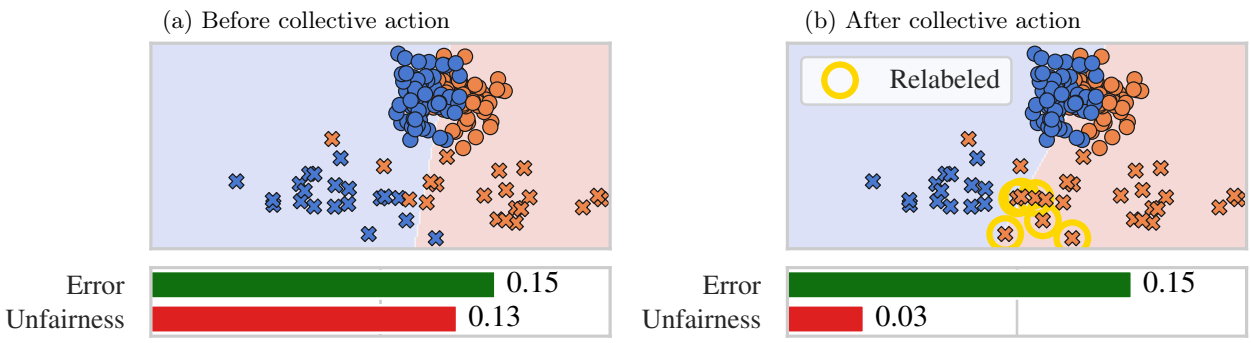

Figure 1: Minority-only collective action can substantially improve fairness. With only 6 label flips, the fairness violation of logistic regression goes down by over 75% with only a negligible increase in prediction error. Circles and crosses represent majority and minority points, respectively.

of majority-group data (Section 4). **(3)** We investigate **fundamental limitations of minority-only collectives** and provide theoretical results showing that **better representations and better counterfactual approximation methods** can improve these algorithms (Section 5).

2 Collective Action for Fairness

To establish the connection between fairness and collective action, Section 2.1 first defines the problem setting and how unfairness can be measured. Then, Section 2.2 describes the theoretical framework of collective action and how it can be utilized to mitigate bias. Finally, Section 2.3 formally relates between collective action to group fairness metrics through counterfactual fairness.

2.1 Group fairness for classification.

We consider a setting in which a firm uses ML to predict a binary label $y \in \{0,1\}$. The firm collects data from its users, forming a dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^m$ denotes user i's feature vector, $a_i \in \{0,1\}$ is a sensitive attribute indicating binary group membership ($a_i = 0$ for the majority group, $a_i = 1$ for the minority), and $y_i \in \{0,1\}$ is the true label. We assume the users are drawn independently and identically distributed (i.i.d.) from a distribution $\mathbb{P}_0$ over $\mathbb{R}^m \times \{0,1\} \times \{0,1\}$. The firm trains a classifier $h : \mathbb{R}^m \to \{0,1\}$ to minimize the prediction error, defined as

$$\text{Error}\,(h) = \mathbb{P}\left[h\,(x) \neq y\right]. \tag{1}$$

To do so, the firm minimizes the empirical error on $\mathcal{D}$ via Empirical Risk Minimization (ERM).

In the group-fairness paradigm, the sensitive attribute $a \in \{0,1\}$ partitions the data into subgroups, and fairness criteria seek to ensure similar outcomes across these groups. Common metrics include statistical parity (SP) [15, 16] and equalized odds (EqOd) [17]. In this work, we focus primarily on violations of EqOd, formally defined as

$$\text{EqOd}\,(h) = \frac{1}{2}\sum_{z=0,1} \left|\mathbb{P}\left[h\,(x) = 1 | a = 1, y = z\right] - \mathbb{P}\left[h\,(x) = 1 | a = 0, y = z\right]\right|, \tag{2}$$

which measures the differences between true positive and false positive rates. Appendix A.1 provides formal definitions and further discussion of these metrics.

ERM-trained models tend to achieve low predictive error, but this often comes at the cost of large fairness violations under SP and EqOd [2, 3, 18, 19]. Despite significant progress in fairness research, most solutions have traditionally focused on *firm-side* solutions: pre-processing the dataset, in-processing modifications to the training algorithm, or post-processing the classifier's predictions. These approaches almost always incur increased error or additional pipeline complexity, discouraging firms to deploy them in practice.

While most prior work has focused on firm-side solutions, this work shifts the focus to *user-side* methods that do not require the firm's participation. Since users generate the training data, they can collectively influence the learned model by strategically modifying their own behavior. For instance, consider a digital platform that recommends content to a user based on classifier predicting engagement labels $y_i \in \{\text{will engage, will not engage}\}$. The classifier, trained on historical user interactions, may unintentionally rely on group membership rather than individual preferences when making recommendations for minority members. In response, users can coordinate to alter their interaction patterns, such as clicking on or avoiding certain items. This collective action affects the dataset in a way that steers the learned classifier toward fairer outcomes. These collectives and their influential abilities in ML are studied as the field of algorithmic collective action [10].

2.2 Algorithmic collective action

In social sciences, *collective action* refers to the coordinated efforts of individuals working together to pursue a shared goal [20]. Hardt et al. [10] adapt this notion to machine learning, proposing that a group of users, termed a collective, can strategically modify their data to align the behavior of a trained classifier h with the collective's goals. In this formulation, the training distribution is a mixture distribution $\mathcal{D} \sim \mathbb{P}_\alpha = \alpha\mathbb{P}^* + (1-\alpha)\mathbb{P}_0$, where $\mathbb{P}^*$ and $\mathbb{P}_0$ are the collective and base distributions, and $\alpha \in [0,1]$ denotes the proportion of the population that belongs to the collective.

Relation to fair representation learning. When users have agency over the training data, one possible form of collective action for fairness is to modify their features to increase correlation with favorable labels. An analogous firm-side approach is fair representation learning (FRL), which learns a transformation from the input space to a representation space such that ERM leads to a classifier that is both accurate and fair [8, 21]. However, a hindrance of FRL in the context of collective action is that the transformation must be applied consistently at inference time, requiring active cooperation from each minority member to transform their features. In contrast, our setting assumes users have control only over the labels and cannot intervene in other parts of the machine learning pipeline.

Erasing a signal. Suppose the collective seeks a classifier that is invariant under a transformation $g : \mathbb{R}^m \to \mathbb{R}^m$ applied to the features. The success of the collective can be quantified as

$$S(\alpha) = \mathbb{P}_0\left[h(g(x)) = h(x)\right], \tag{3}$$

the probability, under the base distribution, that the classifier's prediction remains unchanged after applying g to the features. In words, the collective's goal is to *erase the signal* g: to ensure the classifier behaves identically regardless if the g is applied. Intuitively, if g removes a feature pattern correlated with group membership (e.g., minority vs. majority), then achieving invariance under g promotes fairness by reducing the classifier's dependence on group-identifying information.

To achieve signal erasure, Hardt et al. [10] propose the collective relabels itself with the most likely label under the transformation g. Formally, the strategy is defined as

$$(x, y) \to \left(x, \arg\max_{y' \in \{0,1\}} \mathbb{P}_0\left(y' | g(x)\right)\right). \tag{4}$$

Since this strategy leaves the features unchanged, it is well-suited for settings where the minority is limited to modify only their labels, such as ours. For ϵ-optimal Bayes classifiers (Definition 2), Hardt et al. [10] prove the following lower bound for its success

$$S(\alpha) \geq 1 - \frac{2(1-\alpha)}{\alpha} \cdot \tau - \frac{\epsilon}{(1-\epsilon)\,\alpha}, \tag{5}$$

where $\tau = \mathop{\mathbb{E}}_{x \sim \mathbb{P}_0}\left[\max_{y' \in \{0,1\}} \left|\mathbb{P}_0(y'|x) - \mathbb{P}_0\left(y'|g(x)\right)\right|\right]$ measures the sensitivity of the true label distribution to the transformation g.

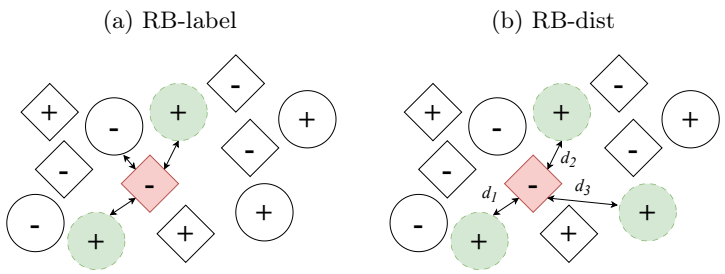

Figure 2: Visualization of KNN scoring methods with k=3. The minority is represented by the squares and the majority by circles, marked with a positive "+" or a negative "−" label. (**a**) *RB-label*: Two of the nearest majority neighbors have a positive label, resulting in the score s=2. (**b**) *RB-dist:* The average distance to the nearest positive majority neighbors results in the score $s = -\left(d_1+d_2+d_3\right)/3$.

Note that the strategy in Equation (4) may require some majority members to relabel themselves with the unfavorable label. Such a change might deter them from participating in the collective action, either because majority members are unwilling to give up their advantage or prefer to maintain the status quo. To avoid this conflict, we restrict the collective to include only minority members. We discuss the implications of this restriction in Section 5.

2.3 Counterfactual fairness

The concept of *counterfactual fairness* (CF) [22, 23, 24] bridges between signal erasure success to group fairness. To introduce this idea, assume that a sample x is generated by a causal model, in which the group membership A is a causal parent. Then a classifier h is counterfactually fair if its predictions are invariant to interventions on the group membership, i.e., $h\left(x\right) = h\left(x_{A\leftarrow a'}\right)$ for any a', where $x_{U\leftarrow u}$ denotes an intervention on a causal parent U of a sample x. In certain causal contexts, CF implies or aligns with group fairness criteria such as SP or EqOd [25]. Therefore, if collective action induces a counterfactually fair classifier, it may also induce a fair classifier under SP or EqOd.

Since our focus is on fairness for the minority group, we relax the original definition of CF [22].

Definition 1. *A classifier h is **minority-focused counterfactually fair** if under any context $X = x$,*

$$\mathbb{P}_0\left(h\left(x_{A\leftarrow a}\right) = y | X = x, A = 1\right) = \mathbb{P}_0\left(h\left(x_{A\leftarrow a'}\right) = y | X = x, A = 1\right), \tag{6}$$

for any value a' attainable by A.

By this condition, changing the group membership of a minority individual, in a counterfactual sense, has no effect on the classifier's prediction. Collective action can theoretically enforce such fairness by applying the erasure strategy from Equation (4) with the counterfactual signal $g\left(x\right) = x_{A\leftarrow 0}$, which replaces a minority individual with its majority-group counterfactual. This collective action aligns the signal erasure success from Equation (3) with minority-focused counterfactual fairness from Definition 1. The following proposition, proved in Appendix B.3, formalizes this alignment.

Proposition 1. *A Bayes classifier trained on $\mathbb{P}_\alpha$ is minority-focused counterfactually fair if and only if the success of a minority collective is $S = 1$.*

This result directly connects between collective action theory to fairness. Thus, perfect success of the collective is equivalent to achieving minority-focused counterfactual fairness.

3 Approximating the Counterfactual Label

This section describes how a minority collective can approximate a signal-erasure strategy to promote fairness in practice. While the theory of signal erasure has been studied before [10, 14], prior work lacks empirical

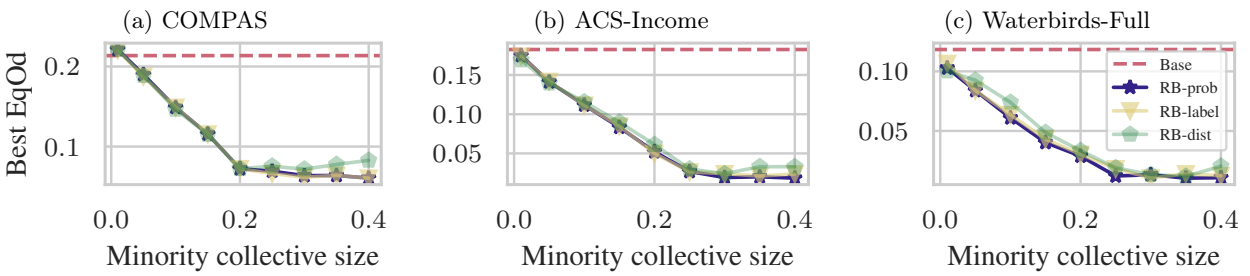

Figure 3: The lowest EqOd violation a collective can achieve greatly improves as the collective size increases, up to a certain point. Each point is a mean of 10 runs, with the standard deviation being smaller than the markers. In all the datasets we experimented on, the lowest EqOd violation converges around $\alpha = 0.3$. Additional results are presented in Figure 10 in the appendix.

evaluation. In this paper, we present the first practical algorithm for signal erasure and provide experimental results in Section 4. As discussed in Section 2.3, a suitable signal to erase is $g(x) = x_{A \leftarrow 0}$, where each collective member relabels themselves according to Equation (4).

However, end-users lack access to the true causal model and cannot compute the counterfactual labels directly. To address this limitation, we propose to assign each collective member i a score s_i, which serves as a proxy for the likelihood that they would receive a favorable label $y = 1$ if they belonged to the majority. Given a budget of M label flips, the collective selects the M members with the highest scores; these individuals flip their labels from $y = 0$ to $y = 1$. The budget M controls the accuracy–fairness tradeoff, where a higher budget typically leads to better fairness, but worse error.

We introduce three model-agnostic scoring functions, each capturing a different notion of similarity to majority users:

1. **Rank by probability (RB-prob):** Train a regressor $f : \mathbb{R}^m \to \mathbb{R}$ on exclusively majority data ($A = 0$) to estimate the probability of receiving a favorable label $\mathbb{P}(Y = 1 | X = x)$. Each collective member i receives a score based on the model's prediction:

$$s_i = f(x_i). \tag{7}$$

2. **Rank by label (RB-label):** For each collective member i, identify the set K_i of their k nearest majority neighbors using Euclidean distance. The score is the number of neighbors with a favorable label:

$$s_i = \sum_{j \in K_i} \mathbf{1}\{y_j = 1\}. \tag{8}$$

3. **Rank by distance (RB-dist):** Restrict the neighbors set K_i to only majority users with favorable label $Y = 1$. The score is the negative mean Euclidean distance to these neighbors:

$$s_i = -\frac{1}{k} \sum_{j \in K_i} \|x_i - x_j\|_2. \tag{9}$$

Intuitively, RB-prob assigns a higher score where a classifier trained solely on majority data predicts a higher likelihood of a favorable label. RB-label scores collective members according to the frequency of favorable labels among their majority neighbors, while RB-dist prioritizes those who are closer majority users with a favorable label. Figure 2 provides visualizations for RB-label and RB-dist.

4 Experimental Results

This section evaluates the performance of our proposed methods under varying settings. We compare the three methods, RB-label, RB-dist, RB-prob, against a random baseline that flips unfavorable labels to

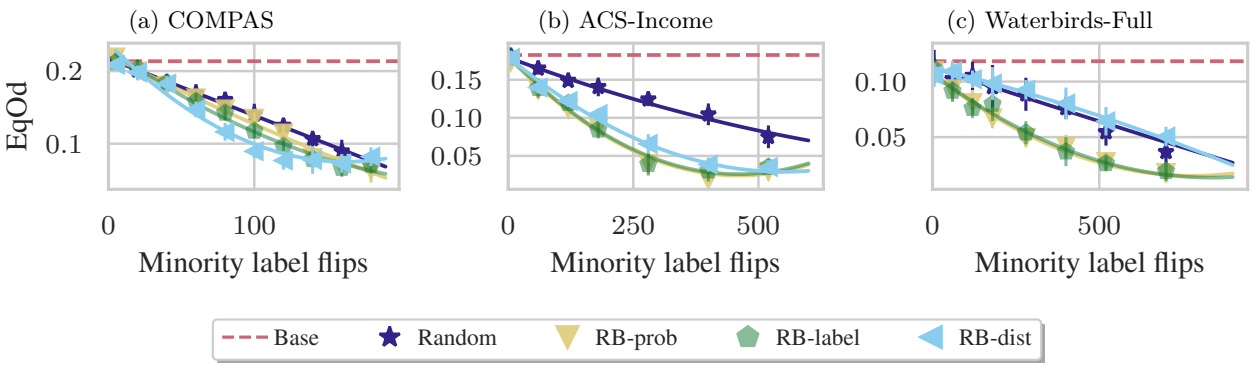

Figure 4: Our proposed methods are consistently more efficient than randomly flipping labels, requiring less label flips to attain the same level of EqOd. Each marker is the mean of 10 random runs with a specific number of label flips. The standard deviation is presented by the error bars. The dashed line shows the mean EqOd for a classifier trained on the dataset without collective action.

favorable for M randomly selected collective members. We conducted experiments on several public datasets: COMPAS [26], Adult [27], HSLS [28], ACS-Income [29] and on two variations of Waterbirds [30]. For Waterbirds, we use features extracted from a pre-trained ResNet-18 (denoted Waterbirds-Full), as well as a version reduced to 85 principal components (denoted Waterbirds-PCA). Details on the datasets and the pre-processing are provided in Appendix C.1.

All reported metrics are computed on a fixed test set, without any collective action, and averaged over 10 independent runs. In each run, we randomly selected a minority collective, which then applied the methods described in Section 3. For the KNN-based methods, we tuned the neighborhood size k using a 15% validation split from the train set, optimizing for EqOd and SP. Finally, we trained a gradient-boosted decision tree on each modified train set.

Importance of collective size While the number of label flips M is the primary factor for balancing between accuracy and fairness, the size of the collective, α, also plays a role. In addition to bounding the possible number of flips, increasing α also expands the candidate pool from which the most effective labels to flip can be selected. To measure this effect, the experiments included a range of α values, each tested with multiple values of M. For each α, we define the best achievable EqOd as the minimum EqOd across all tested values of M. As shown in Figure 3, increasing α improves the best achievable EqOd until saturating around $\alpha = 0.3$. We fix this value for all remaining experiments.

Flipping cost Since each method scores candidates differently, they may also vary in efficiency, that is, the number of label flips required to achieve a given level of fairness. To evaluate efficiency, Figure 4 plots EqOd as a function of number of label flips M, where lower curves indicate more efficient methods. The random baseline consistently yields the worst EqOd across all values of M, highlighting the value of informed relabeling algorithms. However, no single method dominates the others in all settings. While RB-prob and RB-label often outperform the other methods, RB-dist can surpass them in specific cases (e.g., Figure 4a), or perform comparably to the random baseline in others (e.g., Figure 4c).

These results suggest that a well-chosen scoring function enables the collective to achieve a desired level of fairness with fewer label flips, reducing the "cost" of collective action and mitigating the accuracy loss from excessive relabeling.

Interestingly, beyond a certain number of flips, EqOd begins to increase, indicating that excessive flipping can shift unfairness from the minority to the majority. This upturn reflects the fundamental limits of minority collective action for fairness, a point we elaborate on in Section 5.

Partial knowledge of the majority In all previous experiments, we assumed that the collective has full access to the majority data to estimate the counterfactual labels. Here we investigate the performance of

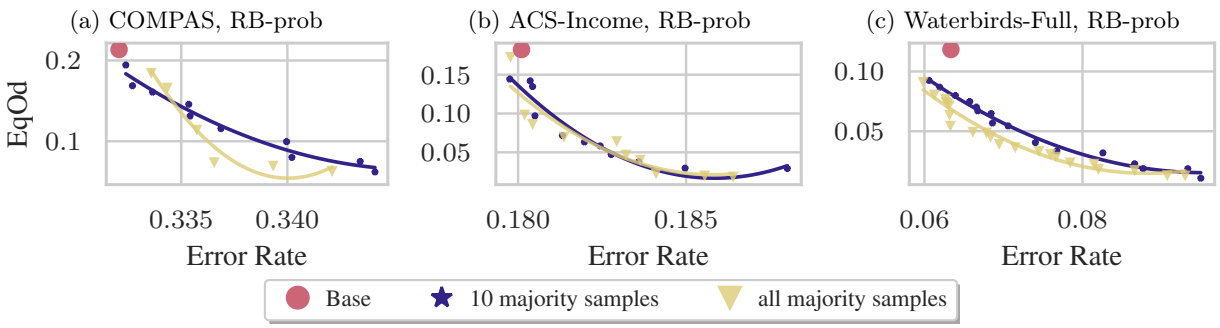

Figure 5: Limiting the knowledge of the collective about the majority does not significantly harm the Pareto front. Each point is the mean of 10 runs and the curves are fitted to guide the eye.

our methods when limiting this knowledge. Figure 5 exhibits that a collective employing RB-prob, when restricted to only 10 majority members, performs similarly as a collective with full knowledge. Additional results for other methods and datasets are provided Figure 14 in the appendix. While the Pareto fronts remain similar, limited majority knowledge can increase the number of required flips. This is evident when comparing to the zero-knowledge scenario, designated as random in Figure 4. This finding implies that the fewer flips the collective is allowed, the more important it is to have access to the majority data.

5 Limitations of Minority Collective Action

Previous work on collective action assumes that the collective is uniformly sampled from the distribution $\mathbb{P}_0$ and that the collective has a perfect oracle for the conditional distribution $\mathbb{P}_0(Y|X)$. Yet, our method restricts collective participation to minority members and approximates this conditional distribution. Those differences introduce limitations to the existing theory, which we analyze and theoretically quantify in this section.

Collective restricted to the minority. As mentioned above, we focus on collectives composed solely of minority members, unlike prior work. This restriction expresses scenarios in which majority members lack incentives to support changes that would benefit the minority, and instead prefer to preserve the status quo. Naturally, this limitation reduces the collective's impact, as demonstrated in the following example.

Consider a binary classification task on the two-dimensional 4-Gaussian mixture model $\mathbb{P}_{4\text{GMM}}$ where each Gaussian belongs to a distinct combination of label and group membership, as illustrated in Figure 6. Each label consists of a large majority subgroup and a significantly smaller minority subgroup. We can then state the following informal result about the EqOd fairness violation of ERM.

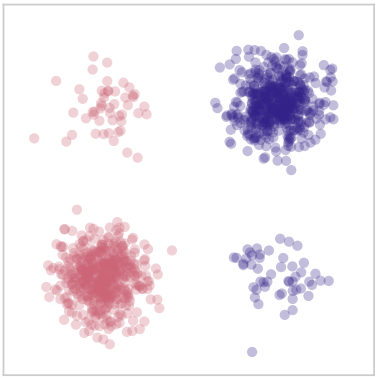

Figure 6: The distribution $\mathbb{P}_{4\text{GMM}}$ used in Proposition 2. The color signifies the label, and the density shows the group membership.

Proposition 2 (Informal)**.** *Consider a dataset sampled from the distribution $\mathbb{P}_{4GMM}$ described above, where every minority point participates in the collective action by flipping all unfavorable labels to favorable. Then, under certain conditions on cluster separation, with high probability, the EqOd of the ERM classifier minimizing the logistic loss will asymptotically approach* 0.5*.*

A formal proposition is provided in Appendix B.1, which holds for a broader family of distributions and can be extended to any dimensionality $\mathbb{R}^d$ using techniques similar to those in Chaudhuri et al. [31]. Although Proposition 2 is not a formal lower bound, it emphasizes an important

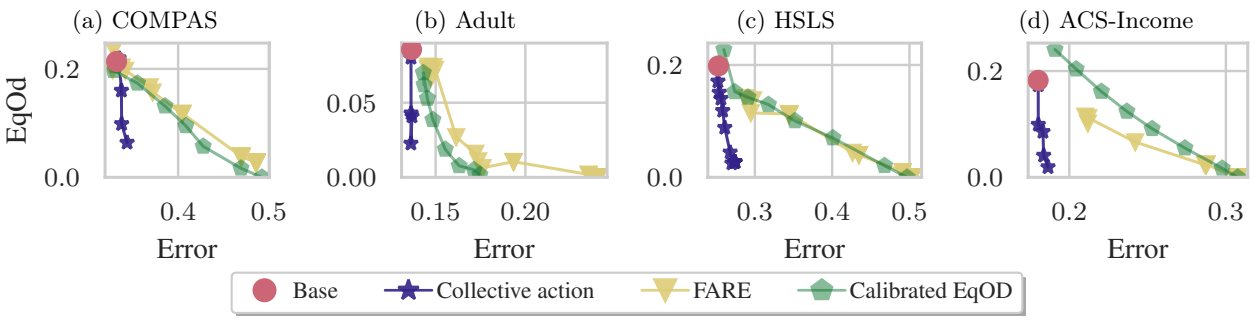

Figure 7: The firm-side pre-processing method FARE [21] and the post-processing method calibrated equalized odds [32] attain 0 EqOd with large error, while RB-prob with $\alpha = 0.3$ (Section 3) has much smaller error and less unfairness than the base classifier, but unable to get 0 EqOd.

limitation: collective action restricted to the minority cannot generally achieve perfect fairness, even under very advantageous conditions involving a maximum-sized collective, a strong strategy, and a complete disregard for accuracy. This limitation stands in contrast to standard firm-side bias mitigation methods, which can, in principle, achieve perfect fairness.

We empirically corroborate the findings of Proposition 2 on real world datasets by examining the fairness–accuracy tradeoff of several fair learning methods. Most of these methods include a hyperparameter that controls this trade-off, yielding a set of pairs (Error, EqOd) as it varies. This set forms a Pareto front, representing the best attainable trade-offs. A Pareto front is said to *dominate* another if it lies entirely to the left (lower error) and below (lower unfairness) of the other.

Figure 7 compares the Pareto fronts of RB-prob, one of our minority collective action methods, with established firm-side methods. We observe that the lowest fairness violation achievable by RB-prob is greater than that of the firm-side approaches. However, the firm-side methods are able to arrive at perfect fairness only at a cost of prohibitively high prediction error. But, inspecting the region where the error is small compared to the base classifier, the fairness of RB-prob is comparable to that of the firm-side methods.

Approximating the class-conditional $\mathbb{P}_0(Y|X)$. In Section 3 we proposed methods to estimate which individuals would receive a different counterfactual label than their original label. However, the success lower bound in Equation (5) assumes perfect knowledge of $\mathbb{P}_0$ and its causal model. To account for approximation error, we model the collective's prediction as the output of algorithm $\mathcal{A}(x) \approx \mathbb{P}_0 \max_y (y|x_{A\leftarrow 0})$ that has an error rate

$$\rho := \mathbb{P}_0 \left(\mathcal{A}(x) \neq \arg\max_{y'} \mathbb{P}_0 [y'|g(x)] \right). \tag{10}$$

Given this definition, we derive the following lower bound on success, proved in Appendix B.2.

Proposition 3. *With algorithm $\mathcal{A}(x)$ with error ρ, the success of the collective action is bounded by*

$$S(\alpha) \geq 1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau - \frac{\epsilon}{(1-\epsilon)(1-2\rho)\alpha}. \tag{11}$$

This bound recovers Equation (5) when $\rho = 0$, but higher values of the error ρ worsen the bound. Next, we show how to use FRL to reduce the error ρ, thereby improving the lower bound.

Impact of feature representations Since the methods RB-label and RB-dist rely on KNN, their performance is sensitive to the choice of distance metric and feature representation. In our main experiments, we used Euclidean distance in the original feature space, which is convenient but could be suboptimal. Here, we explore whether FRL can learn a more suitable representation space for KNN. A *fair representation* maps the data into a space where the group-based bias is removed while preserving informative features. Intuitively, such representations may help RB-label and RB-dist to better estimate the counterfactual labels.

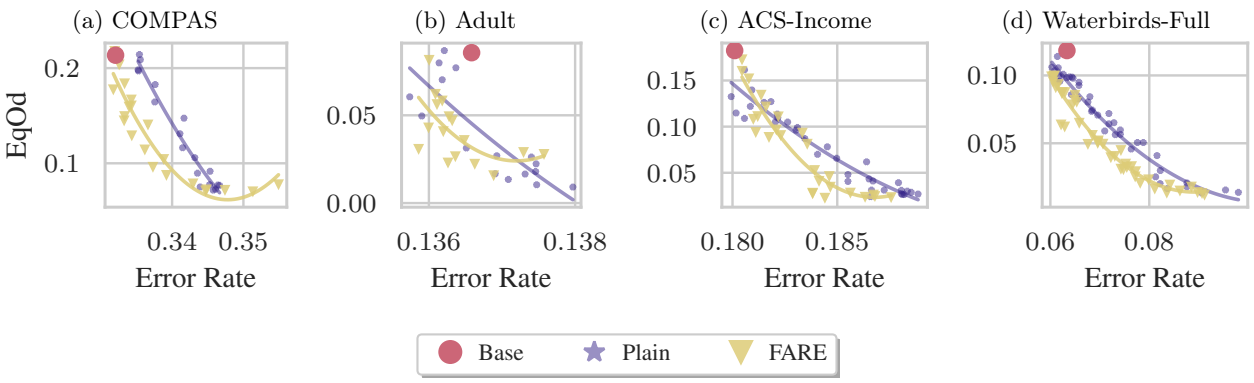

Figure 8: The Pareto fronts for using a fair representation when computing the KNN for RB-dist dominate the Pareto fronts for KNN computed on untransformed features. The blue stars represent the KNN without transforming the data, and the yellow triangles represent the KNN when the data is transformed using FARE [21]. The lines are fitted by a polynomial of degree 2 to guide the eye.

To formalize this intuition, we consider predicting the counterfactual label of minority points using a 1-NN classifier on majority data, i.e., assigning each minority point the label of its nearest neighbor in the majority. In settings where the minority is distributed differently than the majority (e.g., $\mathbb{P}_{4\text{GMM}}$), this task can be challenging. The following informal result compares the error of 1-NN in the original features space to its error in a learned fair representation.

Proposition 4 (Informal)**.** *Let data be drawn from $\mathbb{P}_{4GMM}$, and ρ_{plain} denote the error of a 1-NN classifier that assigns the label of the nearest majority neighbor in the original feature space. Then there exists a fair representation in which a 1-NN classifier achieves error ρ_{FRL} such that, asymptotically with respect to the dataset size, $\rho_{FRL} \leq \rho_{plain}$.*

The formal statement and the proof can be found in Appendix B.4. The result suggests that FRL can reduce the counterfactual label error ρ of RB-label and RB-dist, consequently improving the lower bound of the collective's success according to Proposition 3. Empirically, Figure 8 indicates that applying FARE [21] before the KNN step improves the Pareto front for RB-dist. On the other hand, methods that rely purely on predictive information, such as RB-prob, can perform worse, due to FRL inadvertently removing features predictive of the class label. This behavior, and additional results, are provided in Figure 13 in the appendix.

6 Related work

Optimizing for fairness metrics often comes at the cost of reduced classification accuracy, leading to the well-documented accuracy–fairness trade-off [2, 3, 4, 5]. In response, previous work has proposed fairness interventions at different stages of the ML pipeline: pre-processing methods modify the training data before learning [6, 7, 8, 21], in-processing methods adjust the learning algorithm itself [33, 34, 30, 35], and post-processing methods correct the predictions of a trained (unfair) classifier [17, 36, 37, 38]. A firm can introduce any of these categories into its pipeline, while users, who control only their data can only partially implement pre-processing methods. However, as mentioned in Section 2.2, using feature changing pre-processing methods such fair representation learning [8, 21] demand changing those features during inference time as well.

Still, a couple of pre-processing methods suggest changing only the labels, similarly to our proposed collective action. Luong et al. [7] proposed to compare between the minority KNN and majority KNN and flip the labels according to the difference of positive labels between the two groups of neighbors. This method resembles RB-label, with the difference that RB-label examines only the majority KNN in order to approximate the counterfactual.

Similarly, Kamiran and Calders [6] propose to train a regressor to predict favorable outcome probabilities, and flip the label of minority members with unfavorable labels and high probability according to the regressor to have a favorable label, and similarly flip majority favorable labels to unfavorable. Flipping from both

groups is done to preserve the error of the classifier. Our method RB-prob differs by training the regressor only on the majority to better approximate the counterfactuals.

7 Conclusion

This work demonstrates that user-side methods, specifically minority collective action, can effectively reduce unfairness in machine learning. While much of the existing fairness research focused on firm-side methods, paradoxically these often come at a cost that may not be worth to the firm. This catch emphasizes the importance of studying user-side approaches for bias mitigation. We also note that in general, collective action methods can be exploited by malicious parties seeking self-gain or harming other communities, and it is important to be discussing these limitations and possibly regulate them.

We introduce three practical methods that a collective can easily implement to relabel itself, and show empirically that collective action can considerably reduce unfairness in a variety of datasets, though not completely. Importantly, we also examine the limitations of a minority being composed of only minority members, and how the success is affected by approximating the counterfactual labels.

Overall, this paper shows a practical use case of collective action in the hopes of sparking further research into applications of collective action and user-side methods for social good.

References

[1] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3): 671–732, 2016.

[2] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[3] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Sepehr Dehdashtian, Bashir Sadeghi, and Vishnu Naresh Boddeti. Utility-fairness trade-offs and how to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12037–12046, 2024.

[5] Bashir Sadeghi, Sepehr Dehdashtian, and Vishnu Boddeti. On Characterizing the Trade-off in Invariant Representation Learning. *Transactions on Machine Learning Research*, 2022.

[6] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Control and Communication 2009 2nd International Conference on Computer*, pages 1–6, 2009.

[7] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. K-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510. Association for Computing Machinery, 2011.

[8] Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 2013.

[9] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3384–3393. PMLR, 2018.

[10] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic Collective Action in Machine Learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 12570–12586, 2023.

[11] Omri Ben-Dov, Jake Fawkes, Samira Samadi, and Amartya Sanyal. The Role of Learning Algorithms in Collective Action. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 3443–3461. PMLR, 2024.

[12] Joachim Baumann and Celestine Mendler-Dünner. Algorithmic Collective Action in Recommender Systems: Promoting Songs by Reordering Playlists. In *Advances in Neural Information Processing Systems*, volume 37, pages 119123–119149. Curran Associates, Inc., 2024.

[13] Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünner. Decline now: A combinatorial model for algorithmic collective action. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025.

[14] Etienne Gauthier, Francis Bach, and Michael I. Jordan. Statistical collusion by collectives on learning platforms. In *Forty-Second International Conference on Machine Learning*, 2025.

[15] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. Association for Computing Machinery, 2012.

[17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

[18] Vincent Bardenhagen, Alexandru Tifrea, and Fan Yang. Boosting worst-group accuracy without group annotations. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*. OpenReview, 2021.

[19] Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180, pages 1738–1748. PMLR, 2022.

[20] Mancur Olson. Collective action. In *The Invisible Hand*, pages 61–69. Palgrave Macmillan UK, 1989.

[21] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably Fair Representation Learning with Practical Certificates. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.

[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[23] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. Association for Computing Machinery, 2019.

[24] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1438–1444. International Joint Conferences on Artificial Intelligence Organization, 2019.

[25] Jacy Anthis and Victor Veitch. Causal context connects counterfactual fairness to robust prediction and group fairness. In *Advances in Neural Information Processing Systems*, volume 36, pages 34122–34138. Curran Associates, Inc., 2023.

[26] Jeff Mattu, Julia Larson, Lauren Angwin, and Surya Kirchner. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

[27] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[28] Haewon Jeong, Hao Wang, and Flavio P. Calmon. Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9558–9566, 2022.

[29] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

[30] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *Eighth International Conference on Learning Representations*, 2020.

[31] Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4144–4188. PMLR, 2023.

[32] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[33] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 60–69. PMLR, 2018.

[34] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

[35] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6781–6792, 2021.

[36] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and COMPAS: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*, volume 35, pages 38747–38760. Curran Associates, Inc., 2022.

[37] Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FRAPPÉ: A group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 48321–48343. PMLR, 2024.

[38] André Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024.

8 Appendix

A Preliminaries

A.1 Statistical parity and equalized odds

Among the various ways fairness can be defined in machine learning, group fairness is one of the most studied. Group fairness requires that a model's predictions should not systematically differ between protected groups. One standard measure of this is statistical parity (SP), which captures the difference in the probability of a positive prediction across groups. Formally, it is defined as

$$\mathrm{SP}\left(h\right)=\left|P\left[h\left(x\right)=1|a=1\right]-P\left[h\left(x\right)=1|a=0\right]\right|, \tag{12}$$

where a smaller SP value indicates fairer treatment across groups. However, SP does not account for the ground-truth labels y, and thus optimizing for SP can degrade the overall accuracy. For example, a classifier that always predicts $\hat{y}=1$ will have perfect SP but a high prediction error. Alternatively, a stricter notion called equalized odds (EqOd) [17] requires that both the true positive rate and false positive rate be equal across groups. Here the EqOd difference is defined as

$$\mathrm{EqOd}\left(h\right)=\frac{1}{2}\sum_{z=0,1}\left|P\left[h\left(x\right)=1|a=1,y=z\right]-P\left[h\left(x\right)=1|a=0,y=z\right]\right|. \tag{13}$$

A.2 Suboptimal Bayes classifier

Definition 2 (ϵ-suboptimal classifier)**.** *A classifier $f:\mathcal{X}\to\mathcal{Y}$ is ϵ-suboptimal on a set $\mathcal{X}'\subseteq\mathcal{X}$ under the distribution $\mathbb{P}$ if there exists a $\mathbb{P}'$ with $\mathrm{TV}\left(\mathbb{P}_{Y|X=x},\mathbb{P}'_{Y|X=x}\right)\leq\epsilon$ such that for all $x\in\mathcal{X}'$*

$$f\left(x\right)=\underset{y\in\mathcal{Y}}{\mathrm{argmax}}\,\mathbb{P}'\left(y|x\right).$$

$\mathrm{TV}\left(\cdot,\cdot\right)$ is the total variation distance between two distributions. The definition is discussed more in Hardt et al. [10].

B Theoretical Results and Proofs

B.1 Impossibility of Fairness Under ERM

The following proposition is based on the proof of Thm. 6 from Chaudhuri et al. [31] and assumes the following settings and notations. For notation, $D\left(x\right)$ denotes a distribution with mean x, p is the number of majority samples and m is the number of minority samples with $p\gg m$.

They also assume the following.

Assumption 1 (Concentration Condition, Assumption 2 from Chaudhuri et al. [31])**.** *Suppose $x_1,\ldots,x_n$ are scalars drawn i.i.d from $D(0)$. There exist maps $X_{\max}:\mathbb{Z}_+\times[0,1]\to\mathbb{R}$, $c:\mathbb{Z}_+\times[0,1]\to\mathbb{R}$, and $C:\mathbb{Z}_+\times[0,1]\to\mathbb{R}$ such that for all $n\geq n_0$, for every $\delta\in(0,1)$, with probability $\geq 1-\delta$,*

$$\max_{i\in\{1\cdots,n\}}x_i\in\left[X_{\max}(n,\delta)-c(n,\delta),X_{\max}(n,\delta)+C(n,\delta)\right]$$

Also, $\lim_{n\to\infty}C(n,\delta)=0$, $\lim_{n\to\infty}c(n,\delta)=0$.

They also prove the following.

Lemma 1 (Approximate Maximization Lemma - I, Lemma 14 from Chaudhuri et al. [31])**.** *Let $F(\alpha)=f(\alpha)+g(\alpha)$ where $g(\alpha)=\alpha u+\sqrt{1-\alpha^2}v$, $u,v>0$, and $f(\alpha)$ is an arbitrary function of α that lies in the interval $[-L,U]$. Let α_F be the value of α that maximizes $F(\alpha)$, and let $\alpha_g=\frac{u}{\sqrt{u^2+v^2}}$ be the value of α that maximizes $g(\alpha)$.*

Then, the angle between $(\alpha_F,\sqrt{1-\alpha_F^2})$ and $(\alpha_g,\sqrt{1-\alpha_g^2})$ is at most $\cos^{-1}\left(1-\frac{L+U}{\sqrt{u^2+v^2}}\right)$. Additionally, the maximum value of $F(\alpha)$ is at least $\sqrt{u^2+v^2}-L$.

In their setup, the binary label y and protected attribute a define four groups, each having the same distribution $D(\cdot)$ with different centers. The distributions are defined as $D(y\mu + a\psi)$ where μ, ψ are vectors in $\mathbb{R}^2$ and $\mu = \|\mu\|(1,0)^\intercal$ and $\psi = \|\psi\|(0,1)^\intercal$.

The positive class is denoted as A_μ and the negative class as B_μ. The positive class is split into majority group A_μ^M and minority group A_μ^m, and similarly for the negative class, B_μ^M and A_μ^m. They then define the sets $A_{\mu,\psi}^M = -\psi + A_\mu^M$,$A_{\mu,\psi}^m = \psi + A_\mu^m$,$B_{\mu,\psi}^M = \psi + B_\mu^M$, and $B_{\mu,\psi}^m = -\psi + B_\mu^m$.

In the total absence of the minority group, then an ERM SVM solution will converge to a spurious linear classifier with bias 0 and weight vector $w_{\text{spu}}^* = \frac{\|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}(1,0)^\intercal + \frac{\|\psi\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}(0,1)^\intercal$.

Here, however, here we assume that all the minority samples are applying a relabeling method where each minority sample gets a positive label. As a result, the set $-B_\mu^m$ now has positive labels.

Proposition 5. *Suppose $D(0)$ satisfies the concentration condition in Assumption 1 and $X_{max}(p,\delta,2) - X_{max}(m,\delta,2) \geq 2\|\psi\| + c(p,\delta,2) + C(m,\delta,2)$. If $p \to \infty$, then with probability at least $1-4\delta$, the SVM ERM solution converges to the spurious solution w_{spu}^* with $EqOd(w_{spu}^*) = 0.5$.*

Proof. As Chaudhuri et al. [31] shows, the ERM solution can be written as $w^* = \alpha^*\hat{\mu} + \sigma\beta^*\hat{\psi}$, where

$$\alpha^* = \arg \min_{\alpha\in[-1,1],\sigma\in\{-1,1\}} \sup_{x\in A_\mu \cup -B_\mu} \left(\alpha\hat{\mu} + \sigma\beta\hat{\psi}\right)^\intercal (x-\mu)$$

and $\beta = \sqrt{1-\alpha^2}$.

Case 1: $\sigma = 1$. Here, following Thm 6 and Thm 7 from Chaudhuri et al. [31], splitting the objective to positive and negative sets results in the SVM objective

$$\begin{aligned} F(\alpha) = \min_\alpha \{ & f_3(\alpha) - \alpha\|\mu\| + \beta\|\psi\| \\ & + \max\left(f_1(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| - \beta\|\psi\|\right)\}, \end{aligned}$$

where $f_1(\alpha) = \sup_{x\in A_{\mu,\psi}^M}\left(\alpha\hat{\mu} + \sigma\beta\hat{\psi}\right)^\intercal x$, $f_2(\alpha) = \sup_{x\in A_{\mu,\psi}^m}\left(\alpha\hat{\mu} + \sigma\beta\hat{\psi}\right)^\intercal x$, $f_3(\alpha) = \sup_{x\in -B_{\mu,\psi}^M}\left(\alpha\hat{\mu} + \sigma\beta\hat{\psi}\right)^\intercal x$ and $f_4(\alpha) = \sup_{x\in -B_{\mu,\psi}^m}\left(\alpha\hat{\mu} + \sigma\beta\hat{\psi}\right)^\intercal x$. From conditions on the majority and the minority class, and the Concentration Condition, with probability $1-4\delta$,

$$f_1(\alpha), f_3(\alpha) \in [X_{\max}(p,\delta,2) - c(p,\delta,2), X_{\max}(p,\delta,2) + C(p,\delta,2)]$$

$$f_2(\alpha), f_4(\alpha) \in [X_{\max}(m,\delta,2) - c(m,\delta,2), X_{\max}(m,\delta,2) + C(m,\delta,2)]$$

Observe that from the conditions of the theorem, the first terms will dominate for all values of α , and hence the SVM objective will become

$$F_+(\alpha) = \min_\alpha f_1(\alpha) + f_3(\alpha) - 2\alpha\|\mu\| + 2\beta\|\psi\|.$$

Case 2: $\sigma = -1$. Here the SVM objective becomes

$$\begin{aligned} F(\alpha) = \min_\alpha \{ & f_3(\alpha) - \alpha\|\mu\| - \beta\|\psi\| \\ & + \max\left(f_1(\alpha) - \alpha\|\mu\| - \beta\|\psi\|, f_2(\alpha) - \alpha\|\mu\| + \beta\|\psi\|, f_4(\alpha) - \alpha\|\mu\| + \beta\|\psi\|\right)\}. \end{aligned}$$

This time, from the conditions of the theorem, the first terms will dominate the maximum for all values of α , and hence the objective will become

$$F_-(\alpha) = \min_\alpha f_1(\alpha) + f_3(\alpha) - 2\alpha\|\mu\| - 2\beta\|\psi\|.$$

Note that for all α , $F_-(\alpha) \leq F_+(\alpha)$ and therefore the optimal solution will be for $\sigma = -1$. From Lemma 1, the optimal solution vector $\left(\alpha, \sqrt{1-\alpha^2}\right)$ will be close to the spurious solution vector $\left(\frac{\|\mu\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}, \frac{\|\psi\|}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$,

with the angle being at most $\cos^{-1}\left(1-\frac{C(p,\delta,2)+c(p,\delta,2)}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)$. Under the condition $p\to\infty$, $\cos^{-1}\left(1-\frac{C(p,\delta,2)+c(p,\delta,2)}{\sqrt{\|\mu\|^2+\|\psi\|^2}}\right)\to 0$, resulting with w^*_{spu}.

Note that w^*_{spu} achieves perfect accuracy for the majority with $P_0\left(\hat{y}=1|A=0,y=1\right)=1$ and $P_0\left(\hat{y}=1|A=0,y=0\right)=0$, but goes through the middle of the two minority clusters (where without collective action one cluster is positive and one is negative), resulting in TPR and FPR $P_0\left(\hat{y}=1|A=1,y=1\right)=0.5$ and $P_0\left(\hat{y}=1|A=1,y=0\right)=0.5$. Plugging these values to EqOd results in $\text{EqOd}\left(w^*_{\text{spu}}\right)=0.5$.

□

This result can also be extended to $\mathbb{R}^d$ using techniques similar to those in Chaudhuri et al. [31]. This result also encompasses the 4-Gaussian mixture model $\mathbb{P}_{\text{4GMM}}$ used in Section 5 as a special case, leading to the following.

Proposition 2 (Informal)**.** *Consider a dataset sampled from the distribution $\mathbb{P}_{4GMM}$ described above, where every minority point participates in the collective action by flipping all unfavorable labels to favorable. Then, under certain conditions on cluster separation, with high probability, the EqOd of the ERM classifier minimizing the logistic loss will asymptotically approach* 0.5*.*

B.2 Success Bound With Label Error

The following proof uses Lemma 11 from Hardt et al. [10].

Lemma 2 (Lemma 11 from Hardt et al. [10])**.** *Suppose that P, P' are two distributions such that $\text{TV}(P,P')\le\epsilon$. Take any two events E_1, E_2 measurable under P, P'. If $P(E_1)>P(E_2)+\frac{\epsilon}{1-\epsilon}$, then $P'(E_1)>P'(E_2)$.*

Proposition 3. *With algorithm $\mathcal{A}(x)$ with error ρ, the success of the collective action is bounded by*

$$S(\alpha)\ge 1-\frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau-\frac{\epsilon}{(1-\epsilon)(1-2\rho)\alpha}. \tag{11}$$

Proof. This proof follows closely the proof of Theorem 5 by Hardt et al. [10]. We start under the assumption of an optimal Bayes classifier, setting $\epsilon=0$.

When the new label y' is wrong with probability ρ, then we can think of the collective as being union of two sub-collectives: one with the correct label and one with the incorrect label. In the binary case this can be formulated with correct subcollective P^+ as having label $y'=\arg\max_y P_0(y|g(x))$ and the incorrect subcollective P^- as with label $y'=\arg\min_y P_0(y|g(x))$. Then we can write the train distribution as

$$\begin{aligned} P_\alpha &= \alpha\left(\rho P^- + (1-\rho)P^+\right)+(1-\alpha)P_0 \\ &= \alpha\rho P^- + (1-\rho)\alpha P^+ + (1-\alpha)P_0. \end{aligned} \tag{14}$$

Denote $y^*(x)=\arg\max_y P_0(y|g(x))$, then the probability to get prediction y^* is

$$\begin{aligned} P_\alpha(y^*|x) &= \alpha\rho P^-(y^*|x)+(1-\rho)\alpha P^+(y^*|x)+(1-\alpha)P_0(y^*|x) \\ &= (1-\rho)\alpha+(1-\alpha)P_0(y^*|x), \end{aligned} \tag{15}$$

and the probability to get the prediction $y\neq y^*$ is

$$\begin{aligned} P_\alpha(y|x) &= \alpha\rho P^-(y|x)+(1-\rho)\alpha P^+(y|x)+(1-\alpha)P_0(y|x) \\ &= \alpha\rho+(1-\alpha)P_0(y|x), \end{aligned} \tag{16}$$

where $P^+(y^*|x)=1$, $P^-(y^*|x)=0$, $P^+(y^*|x)=0$, $P^-(y^*|x)=1$ by definition.

A Bayes classifier h returns the most probable label $h(x)=\arg\max_y P(y|x)$. Therefore, a Bayes classifier will output y^* if the probability is greater, which can be written as the condition

$$\begin{aligned} P_\alpha(y^*|x) &> P_\alpha(y|x) \\ (1-\rho)\alpha+(1-\alpha)P_0(y^*|x) &> \alpha\rho+(1-\alpha)P_0(y|x) \\ (1-2\rho)\alpha &> (1-\alpha)\left(P_0(y|x)-P_0(y^*|x)\right). \end{aligned} \tag{17}$$

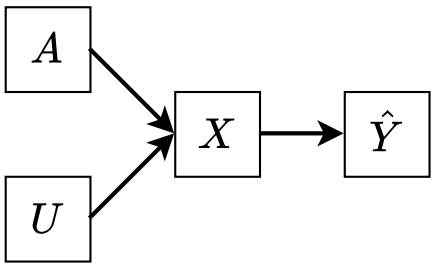

Figure 9: Assumed causal model for data generation and prediction. The group membership A and the other latent variables U are the causal parents of the observable features X. The classifier outputs a predicted label $\hat{Y}$ that depends on the features X.

Let $\tau(x) = \max_y [P_0(y|x) - P_0(y|g(x))]$, then

$$\begin{aligned} P_0(y|x) - P_0(y^*|x) &\leq P_0(y|x) - P_0(y|g(x)) + P_0(y^*|g(x)) - P_0(y^*|x) \\ &\leq 2\tau(x). \end{aligned} \tag{18}$$

With that, the condition in Equation (17) can be written as

$$(1-2\rho)\alpha > 2(1-\alpha)\tau(x). \tag{19}$$

With that, the success can be bounded as

$$\begin{aligned} S &= P_0[f(x) = f(g(x))] \\ &= P_0[f(x) = y^*(x)] \\ &\geq P_0[(1-2\rho)\alpha > 2(1-\alpha)\tau(x)] \\ &= P_0\left[1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau(x) > 0\right] \\ &= \mathbb{E}_{x\sim P_0}\left[\mathbf{1}\left\{1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau(x) > 0\right\}\right] \\ &\geq \mathbb{E}_{x\sim P_0}\left[1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau(x)\right] \\ &= 1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau \end{aligned} \tag{20}$$

With sub-optimality $\epsilon > 0$ A result of Lemma 2 is to write the condition in Equation (19) as

$$(1-2\rho)\alpha > 2(1-\alpha)\tau(x) + \frac{\epsilon}{1-\epsilon}, \tag{21}$$

which by following the same steps as with $\epsilon = 0$ results in the final bound

$$S(\alpha) \geq 1 - \frac{2(1-\alpha)}{(1-2\rho)\alpha}\tau - \frac{\epsilon}{(1-\epsilon)(1-2\rho)\alpha}. \tag{22}$$

□

B.3 Counterfactual Fairness as Success

Proposition 1. *A Bayes classifier trained on $\mathbb{P}_\alpha$ is minority-focused counterfactually fair if and only if the success of a minority collective is $S = 1$.*

Proof. For this proof, we assume the data is generated according to the causal model presented in Figure 9, where the features X are conditioned on the group membership A and other latent causal parent U. The features X are then used by a classifier to compute a predicted label $h(x)\hat{Y}$. In our case, the predicted label is the output of an optimal Bayes classier that predicts the most probable label as $h(x) = \arg\max_y P(y|x)$.

The data distribution is a mixture distribution between the majority distribution $\mathbb{P}_{A=0}$ and the minority distribution $\mathbb{P}_{A=1}$, which is defined as

$$\mathbb{P}_0 = (1-\beta)\,\mathbb{P}_{A=0} + \beta\mathbb{P}_{A=1}, \tag{23}$$

where β is the proportion of the minority in the data.

The collective is employing the signal erasure strategy from Equation (4), where the erased signal is the counterfactual of x if they were a member of the majority group $A = 0$, or formally as

$$g(x) = x_{A\leftarrow 0} \sim \mathbb{P}\left(X_{A\leftarrow 0}\right). \tag{24}$$

The training distribution is a mixture distribution of the data distribution $\mathbb{P}_0$ and the collective distribution $\mathbb{P}^*$, which is defined as

$$\mathbb{P}_\alpha = \alpha\mathbb{P}^* + (1-\alpha)\mathbb{P}_0. \tag{25}$$

We now write the success of the collective (Equation (3)) in terms of the Bayes classifier as

$$\begin{aligned} S &= \mathbb{P}_0\left[h(x) = h(g(x))\right] \\ &= \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha(y|x) = \arg\max_y \mathbb{P}_\alpha(y|g(x))\right]. \end{aligned} \tag{26}$$

To compute this probability, we split it into two cases, conditioning on the group membership A.

When conditioning the success on the majority group $A = 0$, then $g(x) = x$ as the intervention on A, which converts to the majority, does not change the value of A, which is already the majority. This trivially leads to

$$\begin{aligned} S_{A=0} &= \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha(y|x, A=0) = \arg\max_y \mathbb{P}_\alpha(y|g(x), A=0)\right] \\ &= \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha(y|x, A=0) = \arg\max_y \mathbb{P}_\alpha(y|x, A=0)\right] \\ &= 1. \end{aligned} \tag{27}$$

For conditioning the success on the minority, recall that the data is generated according to the causal model in Figure 9 which means that intervention on the group membership A can be passed down to the features X as

$$\mathbb{P}\left(h\left(x_{A\leftarrow 0}\right) = y|X, A=1\right) = \mathbb{P}\left(h(x) = y|X_{A\leftarrow 0}, A=1\right) = \mathbb{P}\left(h(x) = y|g(X), A=1\right). \tag{28}$$

This can be used to write the success conditioned on the minority as

$$\begin{aligned} S_{A=1} &= \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha(y|x, A=1) = \arg\max_y \mathbb{P}_\alpha(y|g(x), A=1)\right] \\ &= \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A=1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A=1\right)\right]. \end{aligned} \tag{29}$$

The first term is rewritten to use the intervention notation even though the intervened variable is unchanged.

As the proportion of the minority is known to be β, the success can be written by combining Equations (27) and (29) using the law of total probability as

$$\begin{aligned} S &= 1 - \beta + \beta\mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A=1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A=1\right)\right] \\ &= 1 - \beta\left(1 - \mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A=1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A=1\right)\right]\right). \end{aligned} \tag{30}$$

This equality can be examined under two scenarios: when the success is perfect $S = 1$ and when the classifier is minority-focused counterfactually fair.

When the success is $S = 1$ If the success of the collective is $S = 1$, then Equation (30) leads to

$$\mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A = 1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A = 1\right)\right] = 1. \tag{31}$$

This means that it is certain that

$$\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A = 1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A = 1\right), \tag{32}$$

Since the label is binary, then it follows that the same applies to using $\arg\min$. Therefore, for all $y \in \{0, 1\}$ we have

$$\mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A = 1\right) = \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A = 1\right), \tag{33}$$

which is the definition of a minority-focused counterfactually fair classifier (Definition 1).

When the classifier is one-sided counterfactually fair If the classifier is one-sided counterfactually fair (Definition 1), then by definition

$$\mathbb{P}_0\left[\arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 1}\right) = y|X, A = 1\right) = \arg\max_y \mathbb{P}_\alpha\left(h\left(x_{A\leftarrow 0}\right) = y|X, A = 1\right)\right] = 1 \tag{34}$$

and plugging that in Equation (30) results in $S = 1$. □

B.4 Label Error With Better Representation

For the following we assume a similar setting as in Appendix B.1, visualised as a 2D distribution in Figure 6. We are given the majority data, and tasked with labeling the minority data. Assume all labels are distributed equally $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1] = \frac{1}{2}$. The minority features $X_{\min}$ are distributed as $X_{\min} \sim \mathcal{N}(y\mu_{\min}, \Sigma_{\min})$ with $X_{\min} \in \mathbb{R}^d$. The label $\hat{y}^{(n)}_{1\text{NN}}$ is predicted according to a 1NN classifier from n majority samples $\mathcal{D}_n = (x_i, y_i)_{i=0}^n$. Majority samples with $y = +1$ are distributed as $X_+ \sim \mathcal{N}(\mu, \Sigma)$, and with $y = +1$ are distributed as $X_- \sim \mathcal{N}(-\mu, \Sigma)$.

Theorem 1. *Assume that $\mu_{min}^\top \Sigma^{-1}\mu > 0$. Further, consider the setting with $\Sigma_{min} = I$, and the minority (i.e. test) distribution introduced above with $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1] = 0.5$ and $X_{min} \sim \mathcal{N}(y\mu_{min}, \Sigma_{min})$.*

Then, there exists a projection $P \in \mathbb{R}^{d\times d}$ such that asymptotically for $n \to \infty$, $err^{rep}_{1NN} < err^{raw}_{1NN}$.

Proof. Consider the projection on the hyperplane perpendicular to w, where $w = \frac{\mu - \mu_{\min}}{2}$. The projection matrix associated with this transformation is $P = I - \frac{ww^\top}{w^\top w}$.

Let us denote the symbols after the projection as $\bar{\mu} := P\mu$, $\bar{\mu}_{\min} := P\mu_{\min}$, $\bar{v} := (P\Sigma P^T)^+\bar{\mu}$ and $\bar{\Sigma}_{\min} := P\Sigma_{\min}P^T$. Here we denoted using A^+ the pseudoinverse of the matrix A. Note that since P is an orthogonal projection matrix, it holds that $PP = P$ and $P^T = P$.

We apply Lemma 3 to obtain closed forms for the asymptotic error of 1NN applied to the initial representation and to the features after the projection P. Namely, using the notation $v := \Sigma^{-1}\mu$ we have:

$$\begin{aligned}
\text{err}_{1\text{NN}} &= \frac{1}{2}\mathbb{P}_{X_{\min}|y=1}[\hat{y}_{1\text{NN}} = -1] + \frac{1}{2}\mathbb{P}_{X_{\min}|y=-1}[\hat{y}_{1\text{NN}} = 1] && (35)\\
&= \frac{1}{2}\left(1 - \Phi\left(\frac{v^\top\mu_{\min}}{\sqrt{v^\top\Sigma_{\min}v}}\right)\right) + \frac{1}{2}\Phi\left(\frac{-v^\top\mu_{\min}}{\sqrt{v^\top\Sigma_{\min}v}}\right) && (36)\\
&= 1 - \Phi\left(\frac{v^\top\mu_{\min}}{\sqrt{v^\top\Sigma_{\min}v}}\right) && (37)\\
&= 1 - \Phi(SNR), && (38)
\end{aligned}$$

where we used the fact that $\Phi(-z) = 1 - \Phi(z)$ and we denote $SNR := \frac{v^\top\mu_{\min}}{\sqrt{v^\top\Sigma_{\min}v}}$.

Similarly, let us denote the SNR corresponding to 1NN applied on the projected representation as follows: $SNR_{\text{proj}} := \frac{\bar{v}^\top \bar{\mu}_{\min}}{\sqrt{\bar{v}^\top \bar{\Sigma}_{\min} \bar{v}}}$.

To show that $\text{err}_{\text{1NN}} > \text{err}^{\text{rep}}_{\text{1NN}}$ it suffices to prove that $SNR < SNR_{\text{proj}}$.

We begin by rewriting the numerator of SNR_{proj}. Since $\mu \in Im(P)$ and because on $Im(P)$ the operators Σ^{-1} and $(P\Sigma P^\top)^+$ represent the same transformation, it follows that:

$$\bar{v} = (P\Sigma P^\top)^+ \bar{\mu} = \Sigma^{-1} \bar{\mu}.$$

Moving on the the denominator of SNR_{proj}, we have that:

$$\begin{aligned}
\bar{v}^\top \bar{\Sigma}_{\min} \bar{v} &= \bar{v}^\top (P \Sigma_{\min} P^\top)^+ \bar{v} \\
&= \bar{v}^\top (P P^\top)^+ \bar{v} \\
&= \bar{v}^\top P^+ \bar{v} \\
&= \bar{v}^\top P \bar{v} \\
&= \bar{v}^\top \bar{v} \\
&= \|\bar{v}\|^2.
\end{aligned}$$

In the second line we used the fact that $\Sigma_{\min} = I$, in the third line we use the identity $P^2 = P$ due to P being a projection matrix, in the forth line we use $P^+ = P$ since P is an orthogonal projection (i.e. P is symmetric) and in the fifth line we use the fact that $\bar{v} \in Im(P)$, and hence, $P\bar{v} = \bar{v}$.

Putting everything together, and using the fact that Σ (and thus, Σ^{-1}) is positive definite (i.e. $x^\top \Sigma^{-1} x > 0, \forall x \in \mathbb{R}^d$) we get that:

$$SNR_{\text{proj}} = \frac{\bar{\mu}^\top \Sigma^{-1} \bar{\mu}}{\|\bar{v}\|^2} > 0 > \frac{\mu^\top \Sigma^{-1} \mu_{\min}}{\|\Sigma^{-1} \mu\|^2} = SNR.$$

□

Lemma 3. *For a unimodal minority distribution $X_{min} \sim \mathcal{N}(\mu_{min}, \Sigma_{min})$ it holds that:*

$$\lim_{n \to \infty} \mathbb{P}_{X_{min}}[\hat{y}^{(n)}_{1NN} = -1] = 1 - \Phi\left(\frac{v^\top \mu_{min}}{\sqrt{v^\top \Sigma_{min} v}}\right),$$

where $v := \mu^\top \Sigma^{-1}$ and Φ is the CDF of a standard Gaussian.

Proof. Let us denote $\hat{y}_{\text{1NN}} := \lim_{n \to \infty} \hat{y}^{(n)}_{\text{1NN}}$ and let p_+ and p_- be the densities of two class-conditional distribution. Notice that the two class conditional training distributions are supported on the entire domain of $\mathbb{R}^d$. Therefore, in the asymptotic regime, the label $\hat{y}_{\text{1NN}}$ at a test point x is given according to the class-conditional distribution that has higher density. Namely, we have:

$$\hat{y}_{\text{1NN}} = \begin{cases} -1 & \text{if } p_+(x) < p_-(x), \\ 1 & \text{otherwise.} \end{cases}$$

Given $X_{\min} \sim \mathcal{N}(\mu_{\min}, \Sigma_{\min})$, we can then write the probability of predicting $\hat{y}_{\text{1NN}} = -1$ as:

$$\mathbb{P}_{X_{\min}}[\hat{y}_{\text{1NN}} = -1] = \mathbb{P}_{X_{\min}}\left[p_+(x) < p_-(x)\right].$$

Using the closed forms for the pdf of a Gaussian, we write the corresponding log-probabilities as follows:

$$\log p_+(x) = -\frac{1}{2}(x - \mu)^\top \Sigma^- 1 (x - \mu) + \text{const.}$$

$$\log p_-(x) = -\frac{1}{2}(x + \mu)^\top \Sigma^- 1 (x + \mu) + \text{const.}$$

Using the fact that log is monotonically increasing and Σ (and by extension Σ^{-1}) is a symmetric matrix, we can write after some simple calculations:

$$\mathbb{P}_{X_{\min}}[\hat{y}_{1\text{NN}} = -1] = \mathbb{P}_{X_{\min}}[\mu^\top \Sigma^{-1} x < 0].$$

Let us denote the random variable $Z := (\mu\Sigma^{-1})X$. Since Z is a linear transformation of Gaussian random variable, it is itself Gaussian and we can write its mean and variance as follows:

$$\mu_Z := v^\top \mu_{\min}, \text{ and } \sigma_Z^2 := v^\top \Sigma_{\min} v, \text{ where } v := \mu^\top \Sigma^{-1}.$$

After this change of variable, we can rewrite the probability of predicting $\hat{y}_{1\text{NN}} = -1$ as:

$$\begin{aligned}
\mathbb{P}_{X_{\min}}[\hat{y}_{1\text{NN}} = -1] &= \mathbb{P}_Z\left[Z < 0\right] \\
&= \Phi\left(\frac{0 - \mathbb{E}[Z]}{\sqrt{\text{Var}[Z]}}\right) \\
&= \Phi\left(\frac{-(\mu^\top \Sigma^{-1})^\top \mu_{\min}}{\sqrt{(\mu^\top \Sigma^{-1})^\top \Sigma_{\min} (\mu^\top \Sigma^{-1})}}\right) \\
&= 1 - \Phi\left(\frac{(\mu^\top \Sigma^{-1})^\top \mu_{\min}}{\sqrt{(\mu^\top \Sigma^{-1})^\top \Sigma_{\min} (\mu^\top \Sigma^{-1})}}\right).
\end{aligned}$$

□

Note that the error from Theorem 1 is defined the same as ρ (Equation (10). This leads to the following.

Proposition 4 (Informal)**.** *Let data be drawn from $\mathbb{P}_{4GMM}$, and ρ_{plain} denote the error of a 1-NN classifier that assigns the label of the nearest majority neighbor in the original feature space. Then there exists a fair representation in which a 1-NN classifier achieves error ρ_{FRL} such that, asymptotically with respect to the dataset size, $\rho_{FRL} \leq \rho_{plain}$.*

C Technical Details

C.1 Datasets

COMPAS The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset contains the data of criminal defendants in Broward county sheriff's office in Florida with the task of predicting the recidivism risk. The label in this dataset represents whether the person re-offended and the sensitive attribute is the race. We follow the same data cleaning and pre-processing as Alghamdi et al. [36].

Adult The Adult dataset [27] contains demographic features of US citizens and is tasked with predicting the income level of an individual. The label represents if the individual has income higher than $50,000 and the sensitive attribute we use is the race. We follow the same data cleaning and pre-processing as Alghamdi et al. [36].

HSLS The High School Longitudinal Study of 2009 (HSLS) [28] contains details of high-school students across the US and the task is to predict the academic success of the students. The label represents the exam score and the sensitive attribute is the race. We follow the same data cleaning and pre-processing as Alghamdi et al. [36].

ACS-Income Ding et al. [29] offer different classification tasks derived by US census data. In our work we used the pre-defined task of predicting level of income denoted as *ACSIncome*, where the data is already pre-processed. The label represents if the individual has income higher than $50,000 and the sensitive attribute is the race.

Waterbirds The waterbirds dataset [30] contains images of landbirds and waterbirds super-imposed on either land or water backgrounds, with the task of classifying the image as of a landbird or a waterbird. The label represents the type of bird, and the sensitive attribute is whether the background is land or water. To obtain the features, we used the output of the penultimate layer of a pre-trained ResNet-18 network from *PyTorch* [1]. We report the results on those features as Waterbirds-Full. We also performed PCA (using *scikit-learn*) and kept the first 85 principal components which retain about 75% of the variance, and report the results of these components as Waterbirds-PCA.

C.2 Training

All classification experiments were trained with *scikit-learn*'s histogram-based gradient boosting classification tree with the default parameters [2]. When there was not a pre-defined test set, we set the train-test split as 80-20 before applying the collective action.

The probabilities for RB-prob were inferred by training *scikit-learn*'s histogram-based gradient boosting classification tree on the majority data with the default parameters, and using its *predict_proba* function. For LFR [8] we used the implementation in *Holistic AI*'s open source library [3] with the default parameters. For FARE [21] we used the official implementation [4] with hyperparameters $\gamma = 0.85$, $k = 200$ and $n = 100$. For all distance computation we used the Euclidean norm ℓ^2-norm as $d(v, u) = \|v - u\|_2 = \sqrt{\sum_i (v_i - u_i)^2}$.

D Additional Results

The following figures include the results of the experiments reported in the main text using all methods on all dataset, both with EqOd (Equation (2)) and SP (Equation (12)) as a measure of unfairness

[1] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html
[2] scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html
[3] https://github.com/holistic-ai/holisticai
[4] https://github.com/eth-sri/fare

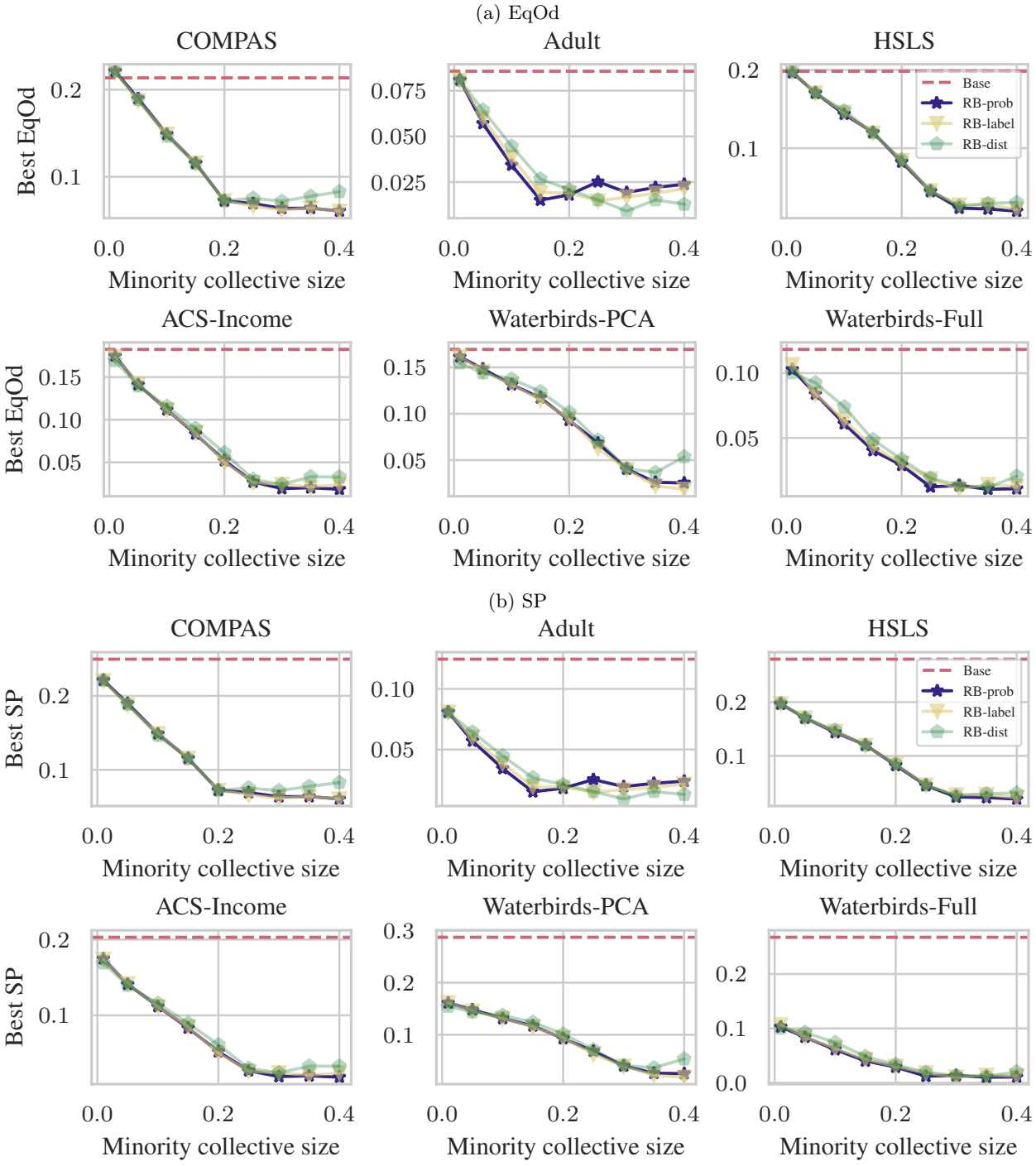

Figure 10: Best achievable fairness with different collective sizes for each of the suggested methods.

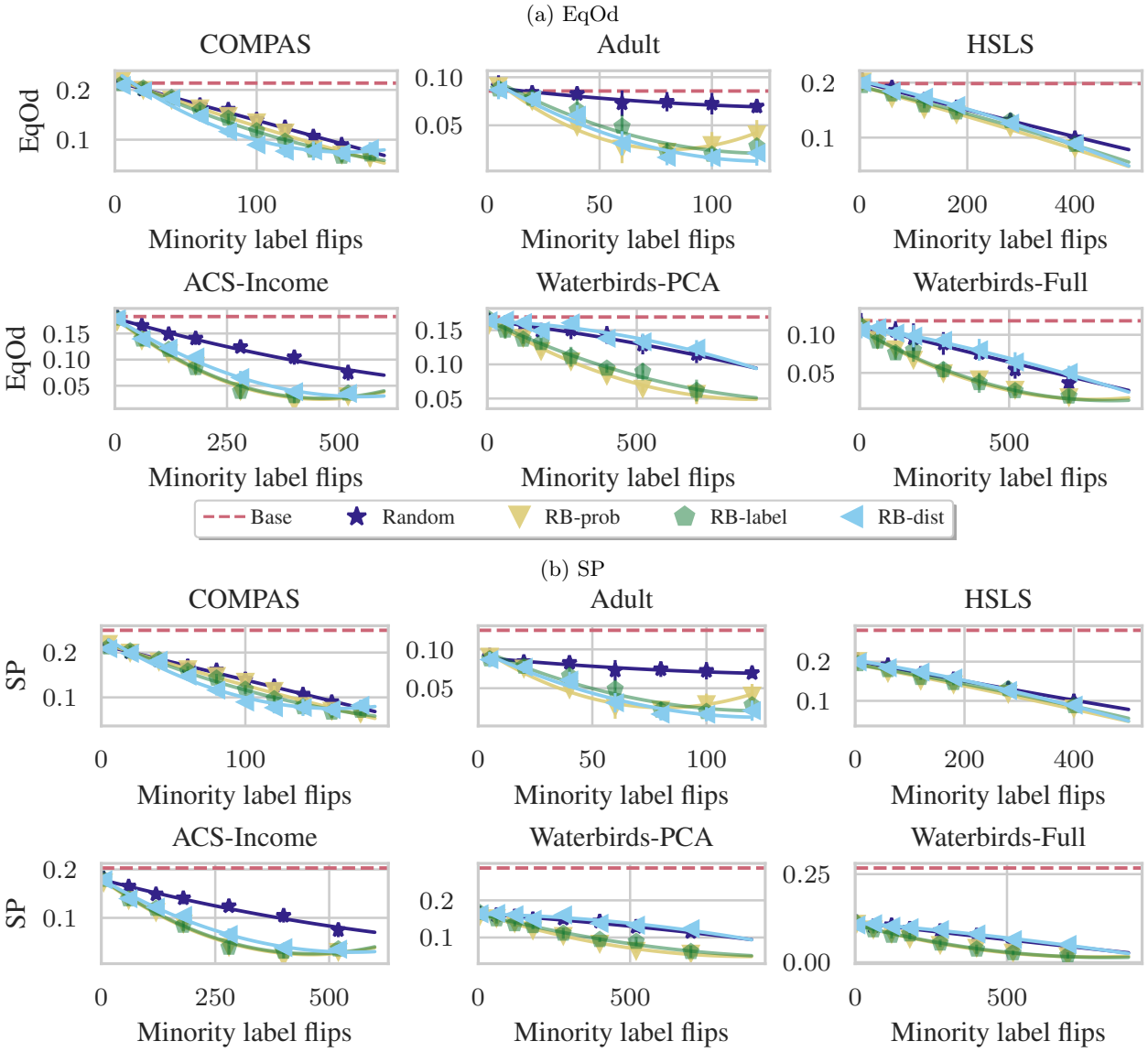

Figure 11: Attained fairness for different numbers of label flips.

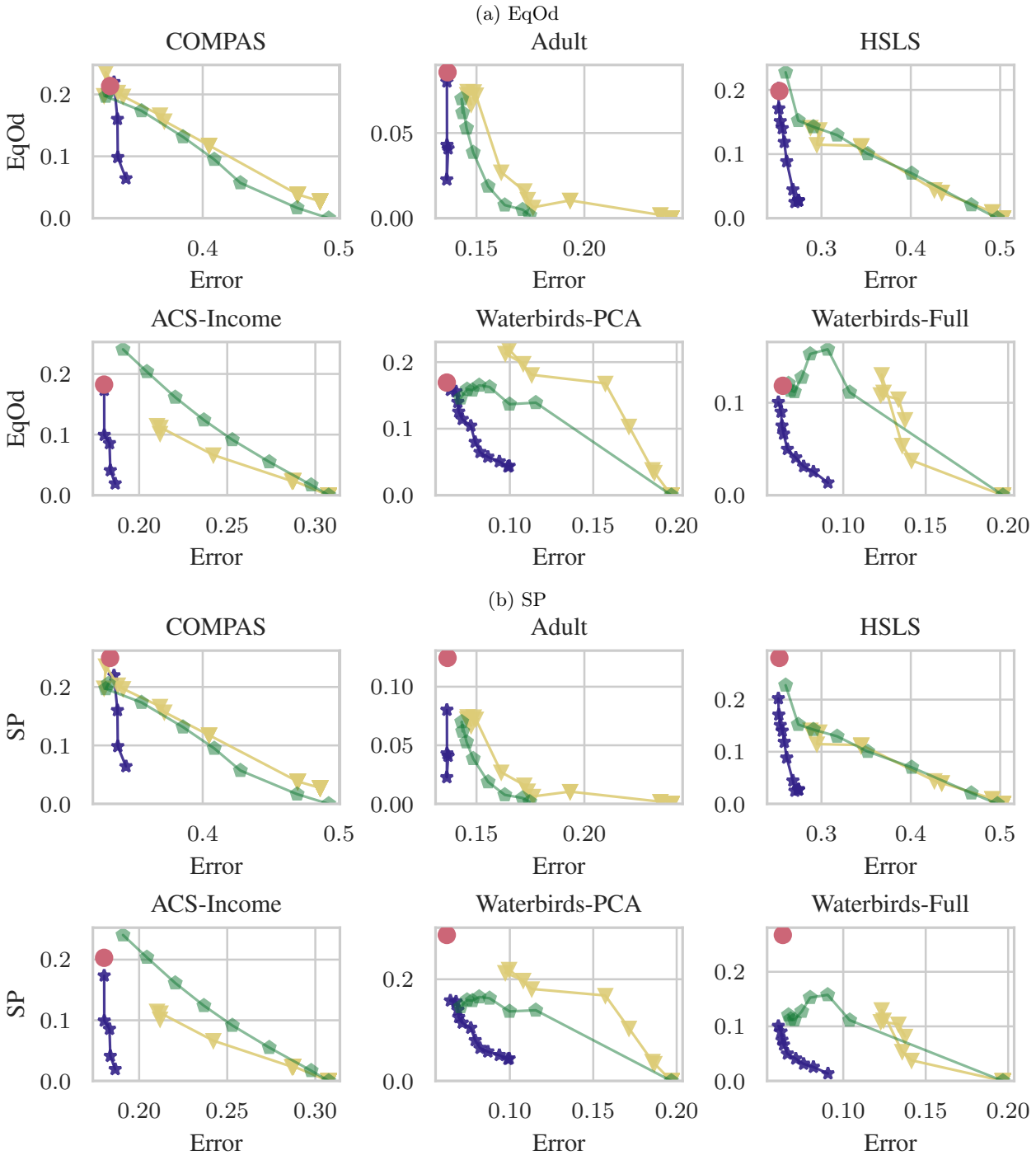

Figure 12: Best achievable fairness when using a classifier without fairness (base), collective action, FARE [21] (pre-processing) and calibrated equalized odds [32] (post-processing).

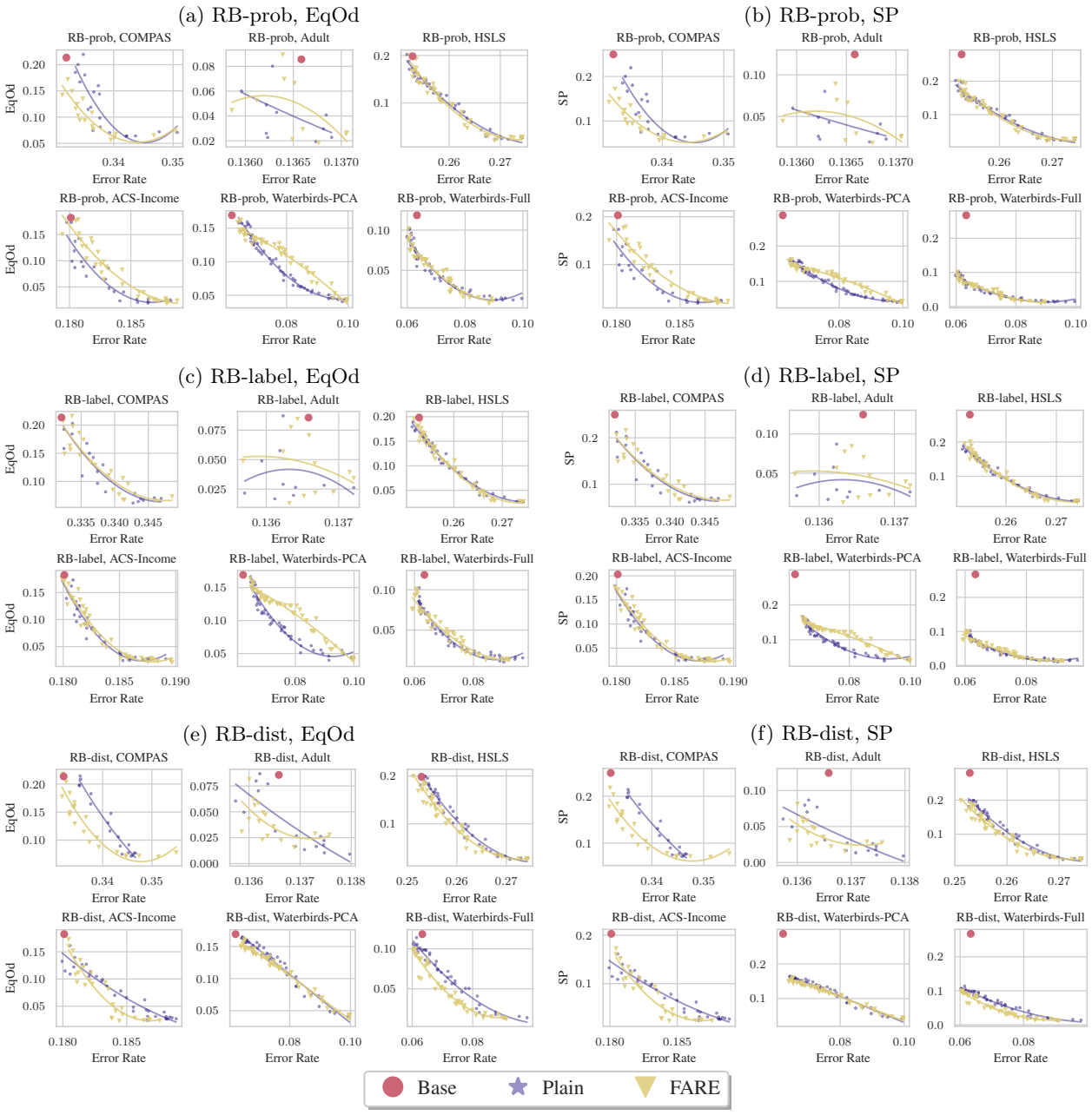

Figure 13: Pareto fronts for each method comparing between using a learned representation and the original features.

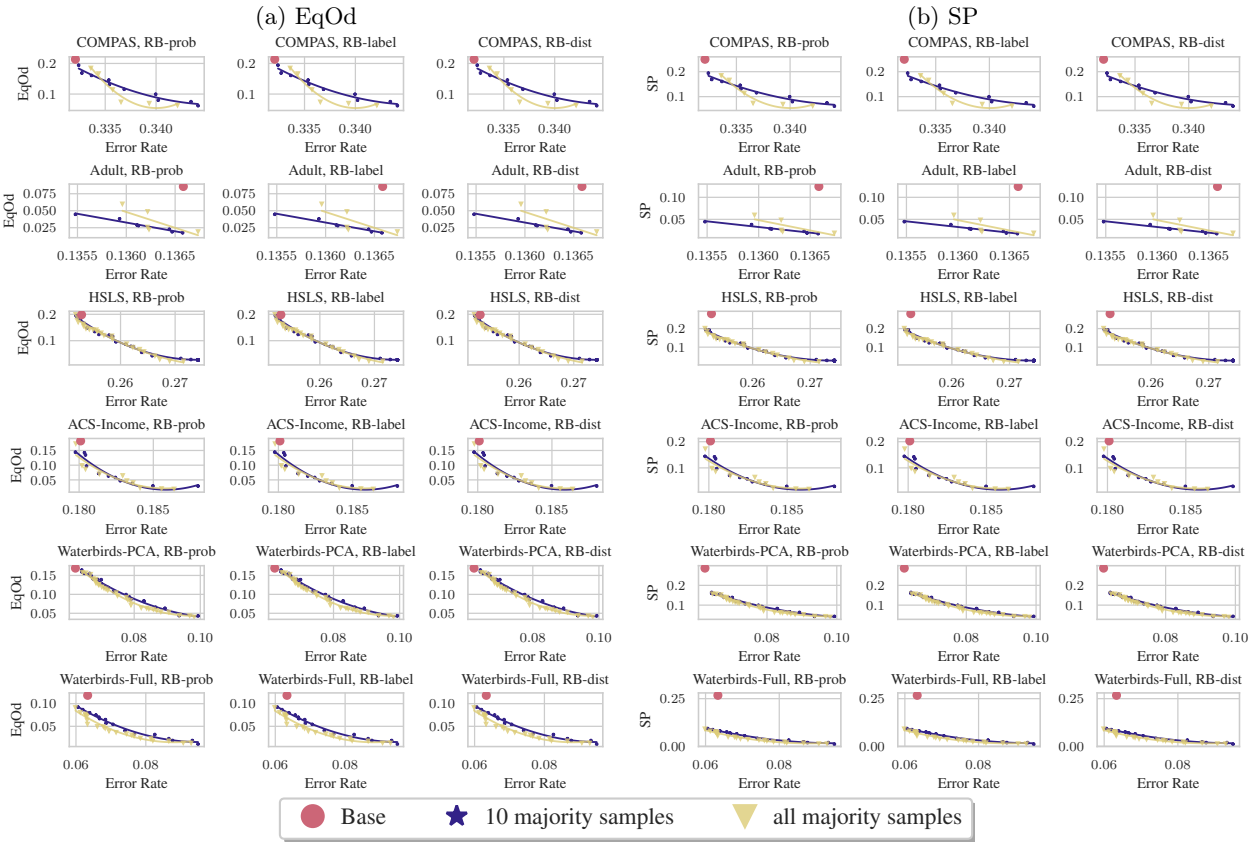

Figure 14: Pareto fronts of our methods when changing the amount of knowledge about the majority.