Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?

Avrim Blum
Toyota Technological Institute at Chicago, 6045 South Kenwood Avenue, Chicago, IL, 60637, USA
avrim@ttic.edu

Kevin Stangl
Toyota Technological Institute at Chicago, 6045 South Kenwood Avenue, Chicago, IL, 60637, USA
kevin@ttic.edu

Abstract

Multiple fairness constraints have been proposed in the literature, motivated by a range of concerns about how demographic groups might be treated unfairly by machine learning classifiers. In this work we consider a different motivation; learning from biased training data. We posit several ways in which training data may be biased, including having a more noisy or negatively biased labeling process on members of a disadvantaged group, or a decreased prevalence of positive or negative examples from the disadvantaged group, or both. Given such biased training data, Empirical Risk Minimization (ERM) may produce a classifier that not only is biased but also has suboptimal accuracy on the true data distribution. We examine the ability of fairness-constrained ERM to correct this problem. In particular, we find that the Equal Opportunity fairness constraint [14] combined with ERM will provably recover the Bayes optimal classifier under a range of bias models. We also consider other recovery methods including re-weighting the training data, Equalized Odds, and Demographic Parity, and Calibration. These theoretical results provide additional motivation for considering fairness interventions even if an actor cares primarily about accuracy.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases fairness in machine learning, equal opportunity, bias, machine learning

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.3

Funding *Avrim Blum*: Supported in part by the National Science Foundation under grants CCF-1815011 and CCF-1733556.
Kevin Stangl: Supported in part by the National Science Foundation under grant CCF-1815011.

Acknowledgements We would like to thank Jon Kleinberg and Manish Raghavan for their helpful and insightful comments on an earlier draft of this manuscript.

1 Introduction

Machine learning (typically supervised learning) systems are automating decisions that affect individuals in sensitive and high stakes domains such as credit scoring [7] and bail assignment [2, 12]. This trend toward greater automation of decisions has produced concerns that learned models may reflect and amplify existing social bias or disparities in the training data. Examples of possible bias in learning systems include the Pro-Publica investigation of COMPAS (an actuarial risk instrument) [2], accuracy disparities in computer vision systems [5], and gender bias in word vectors [4].

In order to address observed disparities in learning systems, an approach that has developed into a significant body of work is to add demographic constraints to the learning problem that encode criteria that a fair classifier ought to satisfy.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 3; pp. 3:1–3:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Multiple constraints have been proposed in the literature [14, 11], each encoding a different type of unfairness one might be concerned about, and there has been substantial work on understanding their relationships to each other, including incompatibilities between the fairness requirements [8, 6, 16, 20].

In this work, we take a different angle on the question of fairness. Rather than argue whether or not these demographic constraints encode intrinsically desirable properties of a classifier, we instead consider their ability to help a learning algorithm to recover from biased training data and to produce a *more accurate* classifier.

In particular, adding a constraint (such as a fairness constraint) to an optimization problem (such as ERM) would typically result in a lower quality solution. However, if the objective being optimized is skewed (e.g., because training data is corrupted or not drawn from the correct distribution) then such constraints might actually help prevent the optimizer from being led astray, and yield a higher quality solution when accuracy is *measured on the true distribution*.

More specifically, we consider a binary classification setting in which data points correspond to individuals, some of whom are members of an advantaged Group A and the rest of whom are members of a disadvantaged Group B. We want to make a decision such as deciding whether to offer a candidate a loan or admission to college. We have access to labeled training data consisting of (x, y) pairs where x is some set of features corresponding to an individual and y is a label we want to predict for new individuals.

The concern is that the training data is potentially biased against Group B in that *the training data systematically misrepresents the true distribution over features and labels in Group B*, while the training data for Group A is drawn from the true distribution for Group A. We consider several natural ways this might occur. One way is that members of the disadvantaged group might show up in the training data at a lower rate than their true prevalence in the population, and worse, *this rate might depend on their true label*.

For instance, if the positive examples of Group B appear at a much lower rate in the training data than the negative examples of Group B (which might occur for cultural reasons or due to other options available to them), then ERM might learn a rule that classifies all or most members of Group B as negative.

A second form of bias in the training data we consider is bias in the labeling process. Human labelers might have inherent biases causing some positive members of Group B in the training data to be mislabeled as negative, which again could cause unconstrained ERM to be more pessimistic than it should be. Alternatively, both processes might occur together. We examine the ability of fairness constraints to help an ERM learning method recover from these problems.

1.1 Summary of Results

Our main result is that ERM subject to the **Equal Opportunity** fairness constraint [14] recovers the true Bayes optimal hypothesis under a wide range of bias models, making it an attractive choice even for decision makers whose overall concern is purely about accuracy on the true data distribution.

In particular, we assume that under the true data distribution, the Bayes optimal classifiers h^*_A and h^*_B classify the same fraction p of their respective populations as positive[1], h^*_A and h^*_B have the same error rate η on their respective populations, and that these errors are uniformly distributed.

[1] $p = P_{\mathcal{D}_A}(h^*_A(x) = 1) = P_{\mathcal{D}_B}(h^*_B(x) = 1)$. We will allow the classifiers to make decisions based on group membership or alternatively assume we have sufficiently rich data to implicitly infer the group attribute.

However, during the training process we do not have access to the true distribution. We only have access to a biased distribution in a way that implicates the distinct social groups and causes the classifier to be overly pessimistic on individuals from Group B.

We prove that, subject to the above conditions on h^*_A and h^*_B, even with substantially corrupted training data either due to the under-representation of positive examples in Group B or a substantial fraction of positive examples in Group B mislabeled as negative, or both, the Equality of Opportunity fairness constraint will enable ERM to learn the Bayes optimal classifier $h^* = (h^*_A, h^*_B)$, subject to a pair of inequalities ensuring that the labels are not too noisy and Group A has large mass.

Expressed another way, this means that *the lowest error classifier on the biased data satisfying Equality of Opportunity is the Bayes optimal classifier on the un-corrupted data.* These results provide additional motivation for considering fairness interventions, and in particular Equality of Opportunity, even if one cares primarily about accuracy.

Other related fairness notions such as Equalized Odds, Demographic Parity, and Calibration do not succeed in recovering the Bayes optimal classifier under such broad conditions. In fact, we show that given data subject to Under-Representation Bias, Calibration can actually *amplify* the effects of the bias, and so can be worse than doing nothing and instead learning with plain ERM (see Section 3.1).

Our results are in the infinite sample limit and we suppress issues of sample complexity[2] in order to focus on the core phenomenon of the data source being unreliable.

1.2 Related Work

This paper is directly motivated by a model of implicit bias in ranking [17]. In that paper, the training data for a hiring process is systematically corrupted against minority candidates and a method to correct this bias increases both the quality of the accepted candidate and the fraction of hired minority candidates. However, that fairness intervention, the Rooney Rule, does not immediately translate to a general learning setting.

Our results avoid triggering the known impossibility results between high accuracy and satisfying fairness criteria [6, 16] by assuming we have equal base rates across groups. This assumption may not be realistic in all settings, however there are settings where bias concerns arise and there is empirical evidence that base rates are equivalent across the relevant demographic groups, e.g. highly differential arrest rates for some alleged crimes that have similar occurrence rates across groups [19, 21].

Within the fairness literature there are several approaches similar to ours. In particular, our concern with positive examples not appearing in the training data is similar in effect to a selective labels problem [18]. [9] uses data augmentation to experimentally improve generalization under selective label bias.

[13, 22] also consider the training and test data distribution gap we experience in our model and posit differing interpretations of fairness constraints under different worldviews. While we do not explicitly use the terminology in these papers, we believe our view of the gap between the true distribution and the training time distribution is aligned with Friedler et al's concept of the gap between the construct space and the observed space.

[2] Our notion of sample complexity is typical. Let S be the biased training data-set and $ERM_{\mathcal{H}}(S) = \hat{h}$. Given $\epsilon, \delta > 0$, $m(\epsilon, \delta)$ samples ensures with probability greater than $1 - \delta$ that $L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \epsilon$.

Our second bias model, Labeling Bias, is similar to [15]. In that paper, the bias phenomenon is that a biased labeler makes poor decisions on the disadvantaged group and intervenes with a reweighting technique, one that is more complex than our Re-Weighting intervention. However, that paper does not consider the interaction of biased labels with different groups appearing in the data at different rates as a function of their labels.

2 Model

In this section we describe our learning model, how bias enters the data-set, and the fairness interventions we consider.

We assume the data lies in some instance space $\mathcal{X}$, such as $\mathcal{X} = \mathbb{R}^d$. There are two demographic groups in the population, Group A and Group B. Their proportions in the population are given by $P(x \in A) = 1 - r$ and $P(x \in B) = r$ for $r \in (0, 1)$. $x \in A$ can be read as individual x in demographic Group A. Group B is the disadvantaged group that suffers from the effects of the bias model.

Assume there is a special coordinate of the feature vector x that denotes group membership. The data distribution is given by $\mathcal{D}$, and is a pair distributions $(\mathcal{D}_A, \mathcal{D}_B)$, with $\mathcal{D}_A$ determining how $x \in A$ is distributed and $\mathcal{D}_B$ determining how $x \in B$ is distributed.

2.1 True Label Generation

Now we describe how the true labels for individuals are generated. Assume there exists a pair of Bayes optimal classifiers $h^* = (h^*_A, h^*_B)$ with $h^*_A, h^*_B \in \mathcal{H} : \mathcal{X} \to \{0, 1\}$.

We assume that the Bayes optimal classifier h^*_B for Group B may be different from the Bayes optimal classifier h^*_A for Group A. If h^*_A was also optimal for Group B, then we can just learn h^* for both Groups A and B using data only from Group A and biased data concerns fade away. Thus we are learning a pair of classifiers, one for each demographic group.

When generating samples, first we draw a data-point x. With probability $1 - r$, $x \sim \mathcal{D}_A$ (and thus $x \in A$) and with probability r, $x \sim \mathcal{D}_B$ (so $x \in B$).

Once we have drawn a data-point x, we model the true labels as being produced as follows; evaluate $h^*(x)$, using the classifier corresponding to the demographic group of x. If $x \in A$, then $h^*(x) = h^*_A(x)$. If $x \in B$, then $h^*(x) = h^*_B(x)$. However, we assume that h^* is not perfect and independently with probability η, the true label of x does not correspond to the prediction $h^*(x)$.

$$y = y(x) = \begin{cases} \neg\, h^*(x) & \text{with probability} \quad \eta \\ h^*(x) & \text{w.p.} \quad 1 - \eta \end{cases}$$

The labels y after this flipping process are the *true labels* of the training data.[3] We assume that $p = P(h^*_A(x) = 1 | x \in A) = P(h^*_B(x) = 1 | x \in B)$. This combined with the assumption that η is the same for classifiers from both groups implies that the two groups have equal base rates (fraction of positive samples) i.e $p(1 - \eta) + (1 - p)\eta$ (un-normalized).

We denote this label model as $(x, y) \sim P_{\mathcal{D},r}(h^*, \eta)$ for a pair of classifiers $h^* = (h^*_A, h^*_B)$ with $h^*_A, h^*_B \in \mathcal{H}$ where $\mathcal{H} : \mathcal{X} \to \{0, 1\}$ is some hypothesis class with finite VC dimension.

[3] Note this label model is equivalent to the Random Classification Noise model [1]. However the key interpretative difference is that in RCN, $h^*(x)$ is the correct label and those that get flipped are noise, but in our case the y are the true labels and h^* is merely the Bayes optimal classifier given the observed features.

2.2 Biased Training Data

Now we consider how bias enters the data-set. Consider the example of hiring where the main failure mode will be a classifier that is too negative on the disadvantaged group. We explore several different bias models to capture potential ways the data-set could become biased.

The first bias model we call **Under-Representation Bias**. In this model, the positive examples from Group B are under-represented in the training data. Specifically, the biased training data is drawn as follows:

1. m examples are sampled from the distribution $\mathcal{D}$. Thus each $x \sim \mathcal{D}$.
2. The label y for each x is generated according to the label process from Section 2.1 with hypothesis $h^* = (h_A^*, h_B^*)$ and η.
3. For each pair (x, y), if $x \in B$ and $y = 1$, then the data-point (x, y) is discarded from our training set independently with probability $1 - \beta$.

Thus we see fewer positive examples from Group B in our training data. β is the probability a positive example from Group B stays in the training data and $1 > \beta > 0$.

If $\eta = 0$, then the positive and negative regions of h^* are strictly disjoint, so if we draw sufficiently many examples, with high probability, we will see enough positive examples in the positive domain of h^* to find a low empirical error classifier that is equivalent to h^*.[4]

In contrast for non-zero η, our label model interacting with the bias model can induce a problematic phenomenon that fools the ERM classifier. For non-zero η there is error even for the Bayes optimal classifier h^* and thus in the region classified as positive by the Bayes optimal classifier h^* there are positive examples mixed with negative examples. The fraction of negative examples is amplified by the bias process.

If β is sufficiently small, there could in fact be more negative examples of Group B than positive examples in the positive region of h_B^*. If this occurs, then the bias model will snap the unconstrained ERM optimal hypothesis (optimal on the biased data) to classifying all individuals from Group B as negatives. This can be observed in Figure 1.

Under-Representation Bias is related to selective labels in [18] since we are learning on a filtered distribution where the filtering process is correlated with the group label. Our model is functionally equivalent to over-representing the negatives of the in the training data, an empirical phenomenon observed in [21].

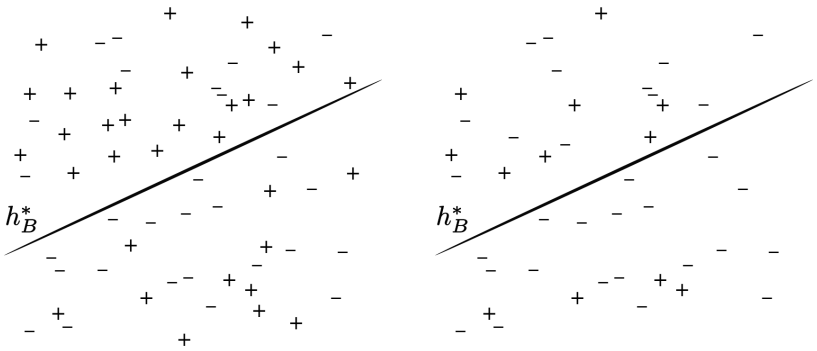
2.3 Alternative Bias Model: Labeling Bias

We now consider a bias model that captures the notion of implicit bias, which we call **Labeling Bias.** In particular, a possible source of bias in machine learning is the label generating process, especially in applications where the sensitive attribute can be inferred by the labeler, consciously or unconsciously. For example, training data for an automated resume scoring system could be based upon the historical scores of resumes created by a biased hiring manager or a committee of experts. This source of labels could then systematically score individuals from Group B as having lower resume scores, an observation noted in randomized real world investigations [3].

Formally, the labeling bias model is:

1. m examples are sampled from the distribution $\mathcal{D}$. Thus each $x \sim \mathcal{D}$.
2. The labels y for each x are generated according to the label process from Section 2.1 with hypothesis $h^* = (h_A^*, h_B^*)$ and η.
3. For each pair (x, y), if $x \in B$ and $y = 1$, then independently with probability ν, the label of this point is flipped to negative.

[4] We would learn with ERM and Uniform Convergence, using the fact that $\mathcal{H}$ has finite VC-dimension.



(a) Un-Corrupted Data

(b) Corrupted Data: Under-Representation Bias

Figure 1 The schematic on the left displays data points with $p = 1/2$, h_B^* as a hyperplane, and $\eta = 1/3$. The schematic on the right displays data drawn from the same distribution subject to the Under-Representation Bias with $\beta_{POS} = 1/3$. Now there are more negative examples than positive examples above the hyperplane so the lowest error hypothesis classifies all examples on the right as negative.

This process is one-sided, so true positives become negatives in the biased training data, so apparent negatives becomes over-represented. We are making a conceptual distinction that the *true* labels (Step 2) are those generated by the original label model and these examples that get flipped by the bias process (Step 3) are not really negative, instead they are just mislabeled.

As ν increases, more and more of the individuals in the minority group appear negative in the training data. Once the number of positive samples is smaller than the number of negative samples above the decision surface h_B^*, then the optimal unconstrained classifier (according to the biased data) is to simply classify all those points as negative.

2.4 Under-Representation Bias and Labeling Bias

We now consider a more general model that combines Under-Representation Bias and Labeling Bias, and moreover we allow either positives or negatives of Group B (or both) to be under-represented. Specifically, we now have *three* parameters: β_{POS}, β_{NEG}, and ν. Given m examples drawn from $P_{\mathcal{D},r}(h^*, \eta)$, we discard each positive example of Group B with probability $1 - \beta_{POS}$ and discard each negative example of Group B with probability $1 - \beta_{NEG}$ to model the Under-Representation Bias. Next, each positive example of Group B is mislabeled as negative with probability ν to model the Labeling Bias. Note that the under-representation comes first: β_{POS} and β_{NEG} represent the probability of *true* positive and *true* negative examples from Group B staying in the data-set, respectively, regardless of whether they have been mislabeled by the agent's labelers.

2.5 Fairness Interventions

Now we introduce several fairness interventions and define a notion of successful recovery from the biased training distribution.

We consider multiple fairness constraints to examine whether the criteria have different behavior in different bias regimes. The fairness constraints we focus on are **Equal Opportunity**, **Equalized Odds**, **Demographic Parity**, and **Calibration**.

▶ **Definition 1.** *Classifier h satisfies* ***Equal Opportunity*** *on data distribution $\mathcal{D}$ [14] if*

$$P_{(x,y)\sim\mathcal{D}}(h(x) = 1|y = 1, x \in A) = P_{(x,y)\sim\mathcal{D}}(h(x) = 1|y = 1, x \in B) \quad (1)$$

This requires that the true positive rate in Group B is the same as the true positive rate in Group A.

Equalized Odds is a similar notion, also introduced in [14]. In addition to requiring Line 1, Equalized Odds also requires that the false positive rates are equal across both groups. Equivalently, we can define **Equalized Odds** as $h \perp A|Y$, meaning that h is independent of the sensitive attribute, conditioned on the true label Y. We also consider **Demographic Parity** $:= P(h(x) = 1|x \in A) = P(h(x) = 1|x \in B)$ [11]. For each of these criteria, the overall training procedure is solving a constrained ERM problem.[5]

An alternative intervention we study **data Re-Weighting**, where we change the training data distribution to correct for the bias process and then do ERM on the new distribution. The overall gist of how the training data becomes biased in our models is that the positive samples from Group B are under-represented in the training data so we can intervene by up-weighting the observed fraction of positives in the training data from Group B to match the fraction of positives from the Group A training data.

In the training process we only have access to samples from the training distribution and thus when using a fairness criterion to select among models *we check the requirement on the biased training data.*

The last fairness intervention we consider is **Calibration**. Calibration [12, 10, 6, 20] requires that when interpreted as probabilities, the same score communicates the same information for individuals from different demographic groups. Specifically, in the bucket of individuals receiving score s, the same fraction in both demographic groups is in fact truly positive. We focus on Calibration for the case of our binary classifier where there are only two scores, e.g. the scores 0 and 1, so in order for classifier $h = (h_A, h_B)$ to satisfy Calibration, the following equalities must hold. [6]

$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 1) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 1)$$
$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 0) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 0)$$

While the other fairness criteria are vigorously debated, Calibration is less contested as an important desiderata of machine learning models. Calibration has been used to defend the epistemic validity of risk prediction instruments [12, 10] and it is claimed that mis-calibrated classifiers may have serious harms and induce undesirable behavior when scores are used by a human actor [20].

Observe that in our model of label generation, the Bayes optimal classifier on the true distribution is the h^* used to generate the labels initially, regardless of the values of η and r. Thus our goal for the learning process is to recover the original optimal classifier h^*, subject to training data from a range of bias models and the true label process with $(x, y) \sim P_{\mathcal{D},r}(h^*, \eta)$. A more effective learning method would recover h^* in a wider range of the model parameters (the parameters that characterize the bias process and the true label process). Accordingly we define **Strong-Recovery**(r, η):

▶ **Definition 2.** *A Fairness Intervention in bias model B satisfies Strong-Recovery(r_0, η_0) if for all $\eta \in [0, \eta_0)$ and all $0 < r < r_0$, when given data corrupted by bias model B, the training procedure recovers the Bayes optimal classifier h^*, given sufficient samples, for all $\beta_{POS}, \beta_{NEG} \in (0, 1]$, $\nu \in [0, 1)$, and $p \in (0, 1]$.*

[5] We do not consider methods for efficiently solving the constrained ERM problem.

[6] If one of the conditioned on events never occurs, such as a classifier that never classifies anyone from Group B as positive, we treat the associated equality as satisfied.

3 Recovery Behavior Across Bias Models

There are two failure modes for learning a fairness constrained classifier that we will need to be concerned with. First, the Bayes optimal hypothesis may not satisfy the fairness constraint evaluated on the biased data. Second, within the set of hypotheses satisfying the fairness constraint, another hypothesis (with higher error on the true distribution) may have lower error than the Bayes optimal classifier h^* on the biased data. We now describe how the multiple fairness interventions provably avoid or fail to avoid these pitfalls in increasingly complex bias models. We defer formal proofs to Section 4.

3.1 Under-Representation Bias

Equal Opportunity and Equalized Odds both perform well in this bias model and avoid both failure modes, subject to an identical constraint on the bias and demographic parameters.

First, from the definition of the Under-Representation Bias model, observe that h^* satisfies both fairness notions on the biased data, so the first failure mode does not occur.

Second, Equal Opportunity intuitively prevents the failure mode where a hypothesis is produced that appears better than h^* on the biased data, such as classifying all examples from Group B as negative, by forcing the two classifiers to classify the same fraction of positive examples as positive. So, if we classify all the examples from Group B as negative, we have to do the same with Group A, inducing large error on the training data from the majority Group A. In particular, so long as the fraction r of total data from Group B is not too large and η is not too close to $1/2$, this will not be a worthwhile trade-off for ERM (saying negative on all samples will not have lower perceived error on the biased data than h^*) and so it will not produce this outcome.

A formal proof of correctness is given in Section 4.1. Specifically, we prove that Equal Opportunity strongly recovers from Under-Representation Bias so long as

$$(1-r)(1-2\eta) + r((1-\eta)\beta - \eta) > 0 \tag{2}$$

Note that this is true for all $\eta < 1/3$ and $r \in (0, 1/2)$, so we have that Equal Opportunity satisfies Strong-Recovery(1/2,1/3) from Under-Representation Bias. Alternatively, we see that if $r = 1/4$ then the inequality simplifies to at least $3/4(1-2\eta) - \eta/4 = 3/4 - (7/4)\eta$ so we have Strong-Recovery$(1/4, 3/7)$. Equalized Odds also recovers in this bias model with the same conditions as Equal Opportunity.

In contrast, Demographic Parity fails to recover h^* even if $\eta = 0$. If $p = 1/2$, $\eta = 0$, and $\beta = 1/2$ and we originally had n samples, then the Bayes optimal classifier does not satisfy Demographic Parity on the biased data since the fraction of samples that will be labelled positive is $\frac{1}{3} \neq \frac{1}{2}$.

Similarly, if we let $\eta \neq 0, \beta < 1$, then in order to match the fraction of positive classifications made by h^*_A, h_B is forced to classify a larger region of the input spaces as positive than h^*_B would in the absence of biased data and so we do not recover h^*_B.

Another way to intervene in the Under-Representation Bias model would just be to re-weight the training data to account for the under-sampling of positives from Group B. If we really know positives from Group B are under-represented, we can change our objective function $min \sum_{i=1}^{m} I(h(x) \neq y)$ by changing each indicator function such that minimizing the sum of indicators measures the loss on the true distribution and not the loss on the biased training distribution.

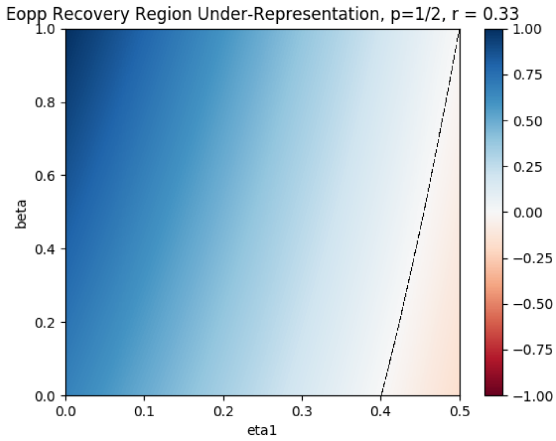


Figure 2 This figure indicates the parameter region such that Equal Opportunity Constrained ERM recovers h^* under the Under-Representation Bias Model and is a visualization of Equation 2. $r = 1/3$ and $p = 1/2$. We label each pair (η, β) with blue if it satisfies the inequality and red otherwise. This plot shows how smaller η means we can recover from lower β. Blue means h^* is recovered. The dashed black line indicates the boundary between recovering h^* and failing to recover h^*.

Define $B^+ = \{x \in B \; s.t. \; y = 1\}$. Then let,

$$I'(h(x), y) = \begin{cases} \frac{1}{\beta} & h(x) \neq 1 \quad and \quad x \in B^+ \\ 0 & h(x) = 1 \quad and \quad x \in B^+ \\ I(h(x) \neq y) & otherwise \end{cases}$$

Then we use this new indicator in the objective function. This new loss function is an unbiased estimator of the true unbiased risk, so uniform convergence on this estimator will suffice to learn h^*. We can infer the value of β from the data for Group A if we know the data from Group B is corrupted by this bias model. One concern with re-weighting in general is that the functional form of the correction is tied to the exact bias model.

As we show in Section 5, Calibration has strange results in this bias model. Specifically, when the bias is such that ERM fails to recover h^* (i.e when $(1-\eta)\beta < \eta$), then the Calibration constraint can only be satisfied by a trivial classifier that assigns all of Group A to one label and all of Group B to the alternative label. For typical parameters, this will result in Group B being given the negative label and Group A will be assigned as all positive. This will not recover h^* and is in fact substantially worse than merely using ERM. Un-constrained ERM would learn badly on Group B but would recover h^*_A for Group A.

When the bias regime is such that $(1-\eta)\beta > \eta$, plain ERM recovers h^*, while enforcing Calibration will lead to excess true error on both demographic groups over the true error of h^*. In particular, satisfying Calibration on the biased data requires intentionally classifying some negative input space from Group A as positive and classifying some positive input space from Group B as negative. These results suggest that Calibration is an actively harmful intervention (for both groups) in our model, when compared to plain ERM, across all model parameters.

In summary, for the Under-Representation Bias model, the fairness interventions Equalized Odds, Equal Opportunity, and Re-Weighting recover h^* under a range of parameters. However, Demographic Parity is inadequate even for $\eta = 0$ and will not recover h^* for non-vacuous bias parameters.

3.2 Labeling Bias

In Section 4, we prove that Equal Opportunity constrained ERM on data biased by the Labeling Bias model also finds the Bayes optimal classifier, under similar parameter conditions to the previous bias model.

Interestingly, in contrast to Under-Representation Bias, *Labeling Bias cannot be corrected* by *Equalized Odds*. The problem is the first failure mode. For example, consider $\eta = 0$ but where $\nu \neq 0$. The Bayes optimal classifier h^*_A for Group A has false positive rate of 0 and true positive rate of 1. However, since $\nu > 0$, there is no classifier for Group B that achieves both of these rates simultaneously. In particular, the only way to classify the negative individuals in the positive region as negative is for the classifier to decrease its true positive rate from 1. Therefore, Equalized Odds rules out usage of h^*_A. This violation holds for $\eta \neq 0$ as well.

In contrast, h^* does satisfy Equal Opportunity on the biased data, and given the conditions in Theorem 3, it will be the lowest error such classifier on the biased data.

Demographic Parity experiences similar limitations as in the Under-Representation Bias model.

The Re-Weighting intervention is to change the weighting of observed positives in the training data for Group B so that we have the same fraction of positives in Group B as in Group A. Define $p_{A,1} :=$ the fraction of positive individuals in Group A and $p_{B,1} :=$ the *observed* fraction of positives in B in the biased data. $p_{A,0}$ and $p_{B,0}$ refer to the observed fraction of negative individuals in Group A and Group B in the biased data.

We need a re-weighting factor Z such that:

$$\begin{aligned}
\frac{p_{A,1}}{p_{A,0}} &= \frac{Zp_{B,1}}{p_{B,0}} \\
\frac{p_{A,1}}{1-p_{A,1}} &= \frac{Zp_{A,1}(1-\nu)}{p_{A,0}+p_{A,1}\nu} = \frac{Zp_{A,1}(1-\nu)}{1-p_{A,1}+p_{A,1}\nu} \\
Z &= \frac{1-p_{A,1}(1-\nu)}{(1-\nu)(1-p_{A,1})}
\end{aligned}$$

We prove in Section 4.2 that this correction factor will lead to the positive region of h^*_B having a higher weight of positive examples than negative examples and simultaneously the negative region of h^*_B having a higher weight of negative examples than positive examples. This causes ERM to learn the optimal hypothesis h^*. We can infer the value of ν by comparing the fraction of positives in Group A and Group B.

In summary, Equal Opportunity and the Re-Weighting Interventions recover well in this bias model (Labeling Bias) while Equalized Odds and Demographic Parity are inadequate.

3.3 Under-Representation Bias and Labeling Bias

In this most general model that combines the two previous models, Re-Weighting the data is now no longer sufficient to recover the true classifier. For example, consider the case where $\eta = 0$ and $p = 1/4$, $\nu = 1/2$ and $\beta_{NEG} = 1/3$ and $\beta_{POS} = 1$. If there were n points originally from group B, then in expectation $3n/4$ were negative and $n/4$ were positive. After the bias process, in expectation there are $n/4$ negatives on the negative side of h^*, and on the positive side of h^* we have $n/8$ correctly labelled positives and what appear to be $n/8$ negative samples.

The Re-Weighting intervention will not do anything in expectation because the overall fractions are still correct; we have $n/2$ total points with one quarter of them labeled positive. ERM is now indifferent between h^* and labeling all samples from Group B as negative. If we just slightly increase the parameter ν and reduce β_{POS} then in expectation ERM will strictly prefer labeling all the samples negatively.

While the Re-Weighting method fails, we prove that Equal Opportunity-constrained ERM recovers the Bayes optimal classifier h^* as long as we satisfy a condition ensuring that Group A has sufficient mass and the signal is not too noisy. As with the previous model, Demographic Parity and Equalized Odds are not satisfied by h^* on minimally biased data and so they will not recover the Bayes optimal classifier.

4 Main Results

We now present our main theorem formally. Define the biased error of a classifier h as its error rate computed on the biased distribution.

▶ **Theorem 3.** *Assume true labels are generated by $P_{\mathcal{D},r}(h^*, \eta)$ corrupted by both Under-Representation bias and Labeling bias with parameters $\beta_{POS}, \beta_{NEG}, \nu$, and assume that*

$$(1-r)(1-2\eta)+r((1-\eta)\beta_{POS}(1-2\nu)-\eta\beta_{NEG})>0 \tag{3}$$

and

$$(1-r)(1-2\eta)+r((1-\eta)\beta_{NEG}-(1-2\nu)\beta_{POS}\eta)>0 \tag{4}$$

Then $h^* = (h^*_A, h^*_B)$ is the lowest biased error classifier satisfying Equality of Opportunity on the biased training distribution and thus h^* is recovered by Equal Opportunity constrained ERM.

Note $\beta_{POS}, \beta_{NEG} \in (0,1]$, $\nu \in [0,1)$, $\eta \in [0,1/2)$, $r \in (0,1)$ and $p \in (0,1]$. Condition 3 refers to Equation 3 and Equation 4.

This case contains our other results as special cases and in the next section we prove our main theorem in this bias model. Note that if Equation 3 is not satisfied then the all-negative hypothesis will have the lowest biased error among hypotheses satisfying Equal Opportunity on the biased training distribution. Similarly, if Equation 4 is not satisfied then the all-positive hypothesis will have the lowest biased error among hypotheses satisfying Equal Opportunity on the biased training distribution. Thus Theorem 3 is tight. To give a feel for the formula in Theorem 3, note that the case of small r is *good* for our intervention, because the advantaged Group A is large enough to pull the classification of the disadvantaged Group B in the right direction. For example, if $r \leq \frac{1}{3}$ then the bounds are satisfied for all $\eta < \frac{1}{4}$ (and if $r \leq \frac{1}{4}$ then the bounds are satisfied for all $\eta < \frac{1}{3}$) for *any* under-representation biases $\beta_{POS}, \beta_{NEG} > 0$ and *any* labeling bias $\nu < 1$.

Thus, Equal Opportunity Strongly Recovers with $(1/4, 1/3)$ and $(1/3, 1/4)$ in the Under-Representation and Labeling Bias model.

Table 1 Summary of recovery behavior of multiple fairness interventions in bias models.

Intervention	Under-Representation	Labeling Bias	Both
Equal Opportunity-ERM	Yes: $(1-r)(1-2\eta)+r((1-\eta)\beta-\eta)>0$	Yes: $(1-r)(1-2\eta)+r((1-\eta)(1-2\nu)-\eta)>0$	Yes: Using Condition 3
Equalized Odds	Yes: $(1-r)(1-2\eta)+r((1-\eta)\beta-\eta)>0$	No	No
Re-weighting Class B:	Yes	Yes	No

Table 1 summarizes the results in the three core interventions and the three core bias models. Demographic Parity is omitted from the table since it cannot recover under the bias models when $\eta = 0$ and thus is inadequate. The contents of each square indicate if recovery is possible in a bias model with an intervention and what constraints need to be satisfied for recovery.

4.1 Proof of Main Theorem

In this section we present the proof of the main result, **Theorem** 3. We want to show that the lowest biased error classifier satisfying Equal Opportunity on the biased data is h^*, given Condition 3.

The first step of the proof is to show that h^* satisfies Equal Opportunity on the biased training data. Note: the lemmas and claims here are all in the Under-Representation Bias combined with Labeling Bias Model, the most general bias model.

▶ **Lemma 4.** *$h^* = (h^*_A, h^*_B)$ satisfies Equal Opportunity on the biased data distribution.*

Proof. First, let's consider the easiest case with $\eta = 0$, $\beta_{POS} = \beta_{NEG} = 1$, and $\nu = 0$. Recall that h^* is the pair of classifiers used to generate the labels. When $\eta = 0$, h^* is a perfect classifier for both groups so Equal Opportunity is trivially satisfied. Now, let's consider arbitrary $0 \leq \eta < 1/2$. Recall that $p = Pr_{\mathcal{D}_A}(h^*_A(x) = 1|x \in A) = Pr_{\mathcal{D}_B}(h^*_B(x) = 1|x \in B)$.

By our assumption that Group A and Group B have equal values of p and η we have

$$\Pr(h^*_A(x) = 1|Y = 1, x \in A) = \frac{p(1-\eta)}{p(1-\eta) + (1-p)\eta} = \Pr(h^*_B(x) = 1|Y = 1, x \in B)$$

Next consider when we have both Under-Representation Bias and Labeling Bias. Recall that $\beta_{POS}, \beta_{NEG} > 0$ is the probability that a positive or negative sample from Group B is *not filtered* out of the training data while $\nu < 1$ is the probability a positive label is flipped and this flipping occurs after the filtering process. Then,

$$\{\text{True Positive Rate on Group A}\} := \Pr(h^*_A(x) = 1|Y = 1, x \in A) =$$
$$\frac{p(1-\eta)}{p(1-\eta) + (1-p)\eta} = \frac{p(1-\eta)\beta_{POS}(1-\nu)}{p(1-\eta)\beta_{POS}(1-\nu) + (1-p)\eta\beta_{POS}(1-\nu)}$$
$$= \Pr(h^*_B(x) = 1|Y = 1, x \in B) := \{\text{True Positive Rate on Group B}\}$$

so Equal Opportunity is still satisfied.

In words, the bias model removes or flips positive points from Group B independent of their location relative to the optimal hypothesis class. Thus positive points throughout the input space are are equally likely to be removed, so the overall probability of true positives being classified as positives is not changed. ◀

Now we describe how a candidate classifier h_B differs from h^*_B. We can describe the difference between the classifiers by noting the regions in the input space that each classifier gives a specific label. This gives rise to four regions of interest with probability mass as follows:

$$p_{1B} = P_{1B}(h_B) := P_{x \in \mathcal{D}_B}(h^*_B(x) = 1 \wedge h_B(x) = 0)$$
$$p_{2B} = P_{2B}(h_B) := P_{x \in \mathcal{D}_B}(h^*_B(x) = 0 \wedge h_B(x) = 1)$$
$$p - p_{1B} = P_{x \in \mathcal{D}_B}(h^*_B(x) = 1 \wedge h_B(x) = 1)$$
$$1 - p - p_{2B} = P_{x \in \mathcal{D}_B}(h^*_B(x) = 0 \wedge h_B(x) = 0)$$

These probabilities are made with reference to the regions in input space *before* the bias process. p_{1B} and p_{2B} are functions of h_B to make explicit that there may be multiple hypotheses with different functional forms that could allocate the same amount of probability mass to parts of the input space where h_B^* and h_B agree on labeling as positive and negative respectively. The partition of probability mass into these regions is easiest to visualize for hyperplanes but will hold with other hypothesis classes. p_{1A} and p_{2A} are defined similarly with respect to h_A^* and $\mathcal{D}_A$. A schematic with hyper-planes is given in Figure 3. To show

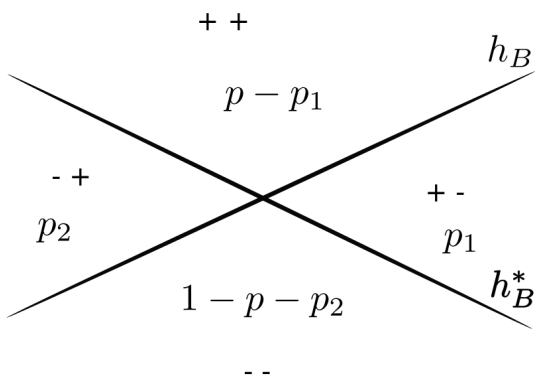


Figure 3 Differences between h_B and h_B^* measured with probabilities in the true data distribution (before the effects of the bias model).

that h^* has the lowest error on the true distribution, we first show how given any pair of classifiers h_A and h_B, which jointly satisfy Equal Opportunity (Equal Opportunity) on the biased distribution, we can transform $\{h_A, h_B\}$ into a pair of classifiers still satisfying Equal Opportunity with at most one non-zero parameter from $\{p_{1B}, p_{2B}\}$, and at most one non-zero parameter from $\{p_{1A}, p_{2A}\}$, while also not increasing biased error.

The final step of our proof argues that out of the family of all hypotheses with (1) at most one non-zero parameter for the hypothesis on Group A, (2) at most one non-zero parameter for the hypothesis on Group B, (3) and jointly satisfying Equal Opportunity on the biased data, h^* has the lowest biased error.

These steps combined imply that h^* is the lowest biased error hypothesis that satisfies Equal Opportunity.

▶ **Lemma 5.** *Given classifiers h_A and h_B which satisfy Equal Opportunity on the biased data, there exist classifiers $h_A^{'}$ and $h_B^{'}$ (not necessarily in $\mathcal{H}$) satisfying*

1. *At most one of $\{P_{1A}(h_A^{'}), P_{2A}(h_A^{'})\}$ is non-zero and at most one of $\{P_{1B}(h_B^{'}), P_{2B}(h_B^{'})\}$ is non-zero.*
2. *$(h_A^{'}, h_B^{'})$ has error at most that of (h_A, h_B) on the biased distribution.*
3. *$h_A^{'}$ and $h_B^{'}$ satisfy Equal Opportunity.*

Proof. We want to exhibit a pair of classifiers with lower biased error that zeros out one of the parameters. We do this by modifying each classifier separately, while keeping the true positive rate on the biased data fixed to ensure we satisfy Equal Opportunity.

First, consider Group A and suppose that $P_{1A}(h_A), P_{2A}(h_A) > 0$ since otherwise we do not need to modify h_A. Imagine holding the true positive rate of h_A constant and shrinking p_{2A} towards zero. As we shrink p_{2A}, we must shrink p_{1A} towards zero in order hold the true positive rate fixed (and thus satisfy Equal Opportunity).

The un-normalized[7] True Positive Rate (constrained by Equal Opportunity) is $(p - p_{1A})(1-\eta) + p_{2A}\eta = p(1-\eta) - p_{1A}(1-\eta) + p_{2A}\eta = (p - p_{1B})(1-\eta) + p_{2B}\eta$. Since the $p(1-\eta)$ term is independent of the classifier h_A, keeping the true positive rate constant is equivalent to keeping $C := -p_{1A}(1-\eta) + p_{2A}\eta$ constant.

Define $f(\Delta) = \Delta\frac{\eta}{1-\eta}$. If $C \leq 0$ then we can shrink p_{2A} to 0 and reduce p_{1A} by $f(p_{2A})$, keeping C constant. If $C \geq 0$ we can instead shrink p_{1A} to 0 and reduce p_{2A} by $f^{-1}(p_{1A})$.

Observe for Group A this process will clearly reduce training error since we are decreasing both p_{1A} and p_{2A} and the error on group A is monotone increasing (and linear) with respect to $p_{1A} + p_{2A}$.

We then separately do this same shrinking process for group B. Now we show the biased error decreases for Group B. For a given amount Δ by which we shrink p_{2B}, the overall biased error change for Group B is $\Delta[\eta\beta_{POS}(1-\nu) - (1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu] + f(\Delta)[\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu - (1-\eta)\beta_{POS}(1-\nu)]$, and simplifies to become

$$= \Delta\eta\beta_{POS}(1-\nu) - f(\Delta)(1-\eta)\beta_{POS}(1-\nu) + \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + f(\Delta)(\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu)$$

The first two terms vanish because of $f(\Delta) = \Delta\frac{\eta}{1-\eta}$.

$$= \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + f(\Delta)(\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu)$$

$$= \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + \Delta\frac{\eta^2}{1-\eta}\beta_{NEG} + \Delta\eta\beta_{POS}\nu$$

$$= \Delta(\frac{\eta^2}{1-\eta}\beta_{NEG} - (1-\eta)\beta_{NEG}) < 0$$

Since this term is negative, we have shown that this modification process decreases error on the biased training data for both Group A and Group B while keeping the true positive rate fixed. $h_A^{'}$ and $h_B^{'}$ are then any functions satisfying these p's (e.g. p_{1A}, p_{2A} etc). ◀

▶ **Lemma 6.** *If h_A and h_B satisfy the Equal Opportunity constraint and each classifier has at most one non-zero parameter, then $p_{1B} = p_{1A}$ and $p_{2B} = p_{2A}$.*

Proof. Recall that the Equal Opportunity constraint requires that these expressions be equal.

$$(p - p_{1A})(1-\eta) + p_{2A}\eta = (p - p_{1B})(1-\eta) + p_{2B}$$
$$p_{2A}\eta - p_{1A}(1-\eta) = p_{2B}\eta - p_{1A}(1-\eta)$$

Then the theorem follows from inspecting the second equality. ◀

This lemma makes explicit that when the classifiers each have only one non-zero parameter and satisfy Equal Opportunity, then the non-zero parameter corresponds to the same region.

▶ **Lemma 7.** *Of hypotheses satisfying ($p_{1A} = p_{1B}$ and $p_{2A} = p_{2B} = 0$) or ($p_{1A} = p_{1B} = 0$ and $p_{2A} = p_{2B}$), if these inequalities hold:*

$$(1-r)(1-2\eta) + r((1-\eta)\beta_{POS}(1-2\nu) - \eta\beta_{NEG}) > 0$$

and

$$(1-r)(1-2\eta) + r((1-\eta)\beta_{NEG} - \eta\beta_{POS}(1-2\nu)) > 0$$

then the lowest biased error classifier satisfying Equal Opportunity on the biased data is $h^* = (h_A^*, h_B^*)$.

[7] The normalization factor for these rates for Group A and Group B is the same so this term can be cancelled.

Proof. First, we sketch the proof informally. Consider three cases which depend on how the bias process affects the unconstrained optimum for Group B on the biased data. In the first case, in the biased data distribution, the region $X^+ := \{x \ s.t. \ h_B^*(x) = 1\}$ has more positive than negative samples in expectation and the region $X^- := \{x \ s.t. \ h_B^*(x) = 0\}$ has more negative than positive samples in expectation. In the second case, there are more positive than negative samples throughout the entire input space in the biased distribution. In the third and final case, there are more negative than positive samples throughout the input space in the biased distribution.

In these three cases, the optimal hypothesis is exactly one of $\{h_B^*, h_B^1, h_B^0\}$, respectively. The second two hypotheses mean labelling all inputs as positive and labelling all inputs as negative, respectively. These three hypotheses correspond to hypotheses with at most one non-zero parameter.

For instance, h_B^1 occurs when $p_{2B} = 1 - p$ and $p_{1B} = 0$. Each of the three hypotheses occur when the one non-zero parameter attains a location on the boundary of its range of values. When p_{2B} is allowed to be non-zero, if instead $p_{2B} = 0$ (and thus it also must be that $p_{1B} = 0$), the hypothesis is equivalent to h_B^*. A similar relationship holds for h^0 and p_1.

In order to show the theorem, we prove that if h^* has lower biased error than $h^1 = (h_A^1, h_B^1)$ and $h^0 = (h_A^0, h_B^0)$ on the biased data distribution, then h^* has the lowest error among all hypotheses with at most one non-zero parameter and satisfying Equal Opportunity.

To see this, consider h_A and h_B with the same non-zero parameter equal to Δ. Then the error of h_A is a linear function of Δ. Similarly, the error of h_B is a linear function of Δ. The overall error of $h = (h_A, h_B)$ is a weighted combination of the error of h^* and the error of h^0 or h^1, so the overall error of h is thus linear in Δ, so the optimal hypothesis parametrized by Δ must occur on the boundaries of the region of Δ, so the optimal hypothesis is one of $\{h^*, h^0, h^1\}$. We then show that the inequalities we assume in the theorem enforce that h^* has strictly lower error than h^0 or h^1. Formally, we enumerate the possible events:

Type	Sign of h^*	Label in Biased Data	Un-Normalized Probability of Event
A	+	+	$R_1 = (1-r)p(1-\eta)$
A	+	-	$R_2 = (1-r)p\eta$
A	-	+	$R_3 = (1-r)(1-p)\eta$
A	-	-	$R_4 = (1-r)(1-p)(1-\eta)$
B	+	+	$R_5 = rp(1-\eta)\beta_{POS}(1-\nu)$
B	+	-	$R_6 = rp[(1-\eta)\beta_{POS}\nu + \eta\beta_{NEG}]$
B	-	+	$R_7 = r(1-p)(\eta\beta_{POS})(1-\nu)$
B	-	-	$R_8 = r(1-p)[(1-\eta)\beta_{NEG} + \eta\beta_{POS}\nu]$

The probabilities on the far right hand side are not normalized. First we show that the $err(h^*) < err(h^1)$. $err(h^*) = R_2 + R_3 + R_6 + R_7$ and $err(h^1) = R_2 + R_4 + R_6 + R_8$, thus $err(h^*) < err(h^1)$ if and only if $R_3 + R_7 < R_4 + R_8$ or thus if

$$(1-r)(1-p)\eta + r(1-p)(\eta\beta_{POS})(1-\nu) \\ < (1-r)(1-p)(1-\eta) + r(1-p)[(1-\eta)\beta_{NEG} + \eta\beta_{POS}\nu]$$

Equivalently,

$$0 < (1-r)(1-2\eta) + r[(1-\eta)\beta_{NEG} - \eta\beta_{POS}(1-2\nu)] \tag{5}$$

Now we consider h^* compared to h^0. Then $err(h^0) = R_1 + R_3 + R_5 + R_7$ Then $err(h^*) < err(h^0)$ if and only if $R_2 + R_6 < R_1 + R_5$.

$$(1-r)p\eta + rp[(1-\eta)\beta_{POS}\nu + \eta\beta_{NEG}] < (1-r)p(1-\eta) + rp(1-\eta)\beta_{POS}(1-\nu)$$

Equivalently,

$$0 < (1-r)(1-2\eta) + r((1-\eta)\beta_{POS}(1-2\nu) - \eta\beta_{NEG}) \tag{6}$$

Thus we have shown that the error of h^* is less than the error of of h^1 and h^0 if and only if both Lines 5 and 6 are true, which we assume in our theorem.

Now we show that we error of $h = (h_A, h_B)$ is linear in Δ. There are two cases depending on what parameter of h is non-zero.

Let h be a hypothesis such that $P_{1A}(h_A) = p_{1B} = \Delta$ and $P_{2A}(h_A) = p_{2B} = 0$ and $\Delta \in [0, p]$.

$$\begin{aligned} err(h) &= R_1\frac{\Delta}{p} + R_2\frac{p-\Delta}{p} + R_3 + R_5\frac{\Delta}{p} + R_6\frac{p-\Delta}{p} + R_7 \\ &= \frac{\Delta}{p}err(h^0) + \frac{p-\Delta}{p}err(h^*) \end{aligned}$$

On the other case let $P_{1A}(h_A) = p_{1B} = 0$ and $P_{2A}(h_A) = p_{2B} = \Delta$ and $\Delta \in [0, 1-p]$.

$$\begin{aligned} err(h) &= R_2 + \frac{1-p-\Delta}{1-p}R_3 + \frac{\Delta}{1-p}R_4 + R_6 + \frac{1-p-\Delta}{1-p}R_7 + \frac{\Delta}{1-p}R_8 \\ &= \frac{\Delta}{1-p}err(h^1) + \frac{1-p-\Delta}{1-p}err(h^*) \end{aligned}$$

Thus the error of h is linear in Δ and boundary values for Δ correspond to the hypotheses in $\{h^*, h^0, h^1\}$. These two arguments show that:

1. Any single parameter h is a weighted sum of (h^* and h^0) or is a weighted sum of (h^* and h^1) and so is linear in Δ. The boundary values of Δ correspond to $\{h^*, h^0, h^1\}$.
2. Since the optimal value of a linear function occurs on the boundaries of its range, the optimal Equal Opportunity classifier with at most one non-zero parameter is one of $\{h^*, h^0, h^1\}$.
3. The inequalities in the theorem statement enforce that h^* has lower biased error than either h^0 or h^1, so h^* has the lowest biased error of any single parameter hypothesis satisfying Equal Opportunity. ◀

If the conditions in the Theorem *do not hold*, then h^* will not have lower error than h^0 and h^1.

4.2 Verification Re-Weighting Recovers from Labeling Bias

The way we intervene by Reweighting is we multiply the loss term for mis-classifying positive examples in Group B by a factor Z such that the weighted fraction of positive examples in biased data for Group B is the same as the overall fraction of positive examples in Group A.

The goal of this reweighting is to ensure that the ratio of positive to negative samples in the positive region of h_B^* is greater than 1 while the ratio is less than 1 in the negative region of h_B^*. Thus the re-weighted probabilities need to simultaneously satisfy:

$$\begin{aligned} \frac{P(y=1|h_B^*(x)=1)}{P(y=0|h_B^*(x)=1)} &= \frac{Z[(1-\eta)(1-\nu)]}{(\eta+(1-\eta)\nu)} > 1 \\ \frac{P(y=1|h_B^*(x)=0)}{P(y=|h_B^*(x)=0)} &= \frac{Z[\eta(1-\nu)]}{((1-\eta)+\eta\nu)} < 1 \end{aligned}$$

The two constraints are equivalent to requiring that:

$$\frac{\eta + (1-\eta)\nu}{(1-\eta)(1-\nu)} < Z < \frac{1-\eta+\eta\nu}{\eta(1-\nu)} \tag{7}$$

Recall from Section 3.2 that $Z = \frac{1-P_{A,1}(1-\nu)}{(1-\nu)(1-P_{A,1})}$

First we show the right hand inequality.

$$\frac{1-p_{A,1}(1-\nu)}{(1-\nu)(1-p_{A,1})} < \frac{1-\eta+\eta\nu}{\eta(1-\nu)}$$
$$0 < \frac{1-\eta+\eta\nu}{\eta} - \frac{1-p_{A,1}(1-\nu)}{(1-p_{A,1})}$$

Observe that both terms are linear in ν. When $\nu = 0$, the inequality becomes $\frac{1-\eta}{\eta} - \frac{1-p_{A,1}}{1-p_{A,1}} = \frac{1-\eta}{\eta} - 1 > 0$. In our bias model $\nu \in [0,1)$, but if $\nu = 1$, the inequality becomes $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$. Thus Equation 7 holds if both $\frac{1-\eta}{\eta} - 1 > 0$ and $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$.

$\frac{1-\eta}{\eta} - 1 > 0$ is clearly true because $0 < \eta < 1/2$.

To see that $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$, note that this is equivalent to $\eta < 1 - p_{A,1}$, where the right-hand-side is the overall fraction of negative examples in A. This is clearly true because the positive region of h^*_A has exactly an η fraction of negatives, and the negative region of h^*_A has a $1-\eta > \eta$ fraction of negatives.

Now we show the left hand inequality in Equation 7.

$$\frac{\eta + (1-\eta)\nu}{(1-\eta)(1-\nu)} < \frac{1-P_{A,1}(1-\nu)}{(1-\nu)(1-P_{A,1})}$$
$$\frac{\eta + (1-\eta)\nu}{(1-\eta)} < \frac{1-P_{A,1}(1-\nu)}{1-P_{A,1}}$$
$$0 < \frac{1-P_{A,1}(1-\nu)}{(1-P_{A,1})} - \frac{\eta + (1-\eta)\nu}{(1-\eta)} \tag{8}$$

We follow a similar linearity argument to above. For $\nu = 1$, Equation 8 becomes $\frac{1}{1-p_{A,1}} - \frac{1}{1-\eta} > 0$. This holds if $1 - p_{A,1} < 1 - \eta \iff \eta < p_{A,1}$. This is clearly true because the negative region of h^*_A has exactly an η fraction of positives, and the positive region of h^*_A has a $1-\eta > \eta$ fraction of positives. For $\nu = 0$, Equation 8 becomes $1 - \frac{\eta}{1-\eta} > 0$ which holds since $0 < \eta < 1/2$.

5 Calibration Results

▶ **Theorem 8.** *Assume the training data is corrupted by Under-Representation Bias with parameter $\beta < 1$. For any such β, h^* does not satisfy Calibration on the biased data and thus Calibration constrained ERM will return a hypothesis that has strictly worse true error than the true error of h^*. This occurs even when $(1-\eta)\beta > \eta$, i.e. in the bias regime such that plain ERM on the biased data would recover h^*.*

Moreover, if bias is such that $(1-\eta)\beta < \eta$ and thus ERM on the biased data will not recover h^*, then the unique ERM solution that satisfies Calibration on the biased data is a trivial classifier, meaning that all individuals from Group A receive one label (the positive label) and all individuals from Group B receive the opposite label.

Proof. Recall that Calibration of hypothesis $h = (h_A, h_B)$ requires that both Eq. 9 and 10 hold simultaneously.

$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 1) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 1) \tag{9}$$
$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 0) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 0) \tag{10}$$

We assume that if one of the terms is vacuous in the Calibration constraints , then that constraint is still satisfied. In other words, if one bin is non-empty for one group while the corresponding bin for the other group is empty, we assume that bin satisfies Calibration. Due to the effects of the bias model positive samples from Group B appear in the training data with lowered frequency and so the equalities in Equations 9 and 10 become:

$$P_{x\sim A}(y = 1|h^*_A(x) = 1) > P_{x\sim B}(y = 1|h^*_B(x) = 1) \tag{11}$$
$$P_{x\sim A}(y = 1|h^*_A(x) = 0) > P_{x\sim B}(y = 1|h^*_B(x) = 0) \tag{12}$$

Thus $h^* = (h^*_A, h^*_B)$ violates calibration for any $\beta < 1$ and any other hypothesis satisfying calibration will have strictly greater error on the true data distribution. Intuitively, for h to be Calibrated it will need to reduce the left-hand side of Equation 11 because it cannot increase the right-hand side and will have to increase the right-hand side of Equation 12 because it cannot decrease the left-hand side. As a result, its true error will be strictly larger than that of h^*.

Now, consider $(1 - \eta)\beta < \eta$. In this case, plain ERM will not recover h^*. With this amount of bias, then:

$$\begin{aligned} P_{x\sim A}(y = 1|h^*_A(x) = 1) &> P_{x\sim A}(y = 1|h^*_A(x) = 0) \\ &> P_{x\sim B}(y = 1|h^*_B(x) = 1) > P_{x\sim B}(y = 1|h^*_B(x) = 10) \end{aligned}$$

Satisfying Calibration with non-trivial classifiers requires achieving an equality with one side being a non-negative combination of the first two probabilities, and the other side being a non-negative combination of the second two probabilities. Since these inequalities are all strict, this is clearly not possible, so the only way to satisfy calibration is to use a trivial classifier that assigns all of Group A to one label, and all of Group B to the other label.[8] ◀

6 Conclusion

In this paper we have shown that Equal Opportunity constrained ERM will recover from several forms of training data bias, including Under-Representation Bias (where positive and/or negative examples of the disadvantaged group show up in the training data at a lower rate than their true prevalence in the population) and Labeling Bias (where each positive example from the disadvantaged group is mislabeled as negative with probability $\nu \in (0, 1)$), in a clean model where the Bayes optimal classifiers h^*_A, h^*_B satisfy most fairness constraints on the *true* distribution and the errors of h^*_A, h^*_B are uniformly distributed. The high-level message of this paper is that fairness interventions need not be in competition with accuracy and may improve classification accuracy if training data is unrepresentative or biased; however these results will be connected to the true data distributions and features of

[8] Which trivial classifier is selected by ERM will depend on p and r. If $1 - r > r$ and $p > 1/2$, then Group A will be all positive and Group B all negative. While if $1 - r > r$ and $p < 1/2$, then then Group A will be all positive and Group B all negative.

the biased data-generation process. It would be interesting to consider other ways in which training data could be biased, and other assumptions on the optimal classifiers, to determine what kinds of interventions might be most appropriate for different biased-data scenarios.

References

1. Dana Angluin and Philip Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, April 1988. doi:10.1007/BF00116829.
2. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.
3. Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004.
4. Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.
5. Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.
6. Alexandra Chouldechova. Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.
7. Danielle Keats Citron and Frank Pasquale. The Scored Society: Due Process for Automated Predictions. *Wash. L. Rev.*, 89:1, 2014.
8. Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.
9. Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint*, 2018. arXiv:1807.00905.
10. William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.
11. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness Through Awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012. doi:10.1145/2090236.2090255.
12. Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80:38, 2016.
13. Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016. arXiv:1609.07236.
14. Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.
15. Heinrich Jiang and Ofir Nachum. Identifying and Correcting Label Bias in Machine Learning. *CoRR*, abs/1901.04966, 2019. arXiv:1901.04966.
16. Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017. doi:10.4230/LIPIcs.ITCS.2017.43.
17. Jon M. Kleinberg and Manish Raghavan. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 33:1–33:17, 2018. doi:10.4230/LIPIcs.ITCS.2018.33.

18 Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284. ACM, 2017.

19 Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

20 Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

21 Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online, Forthcoming*, 2019.

22 Samuel Yeom and Michael Carl Tschantz. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. *arXiv preprint*, 2018. `arXiv:1808.08619`.