

# Balancing Fairness and Accuracy in Decision Making

[Fairness Constraints: Mechanisms for Fair Classification, Zafar et.al 2015](#)

**Article by Pratyush Garg**

Educated decision making has traditionally been something that humans have claimed as one of their most unique characteristics. In recent years, however, artificial intelligence and machine learning has challenged this monopoly by showing an ability to analyze historical data and leverage inherent patterns and information for prediction and making decisions. These new tools have subsequently been applied in almost every domain imaginable -- and continue to find new avenues everyday.

While these advances are clear to see, recent attention has also been drawn to the risk that algorithms can raise concerns relating to bias. In domains where the impact of such a decision is considerable (criminal recidivism prediction, hiring decisions, etc), the presence of algorithmic bias strikes in a two-pronged manner -- first, by virtue of being algorithmic/mathematical in nature, it serves to put practitioners at ease with the idea that “math is inherently fair” and second, by actually propagating and amplifying historical injustices. This recognition has therefore led to a lot of research focusing on identifying and mitigating sources of bias in the algorithmic decision making pipeline.

For the sake of clarity of the text, we will use the particular context of algorithmic hiring, as the domain of decision making under consideration. We will also assume that the employer is using some sort of a classifier to separate candidates into the “hireable” and “rejected” classes. Our problem then becomes to understand the questions of accuracy and fairness of the outcomes.

## **BIAS IN LAW**

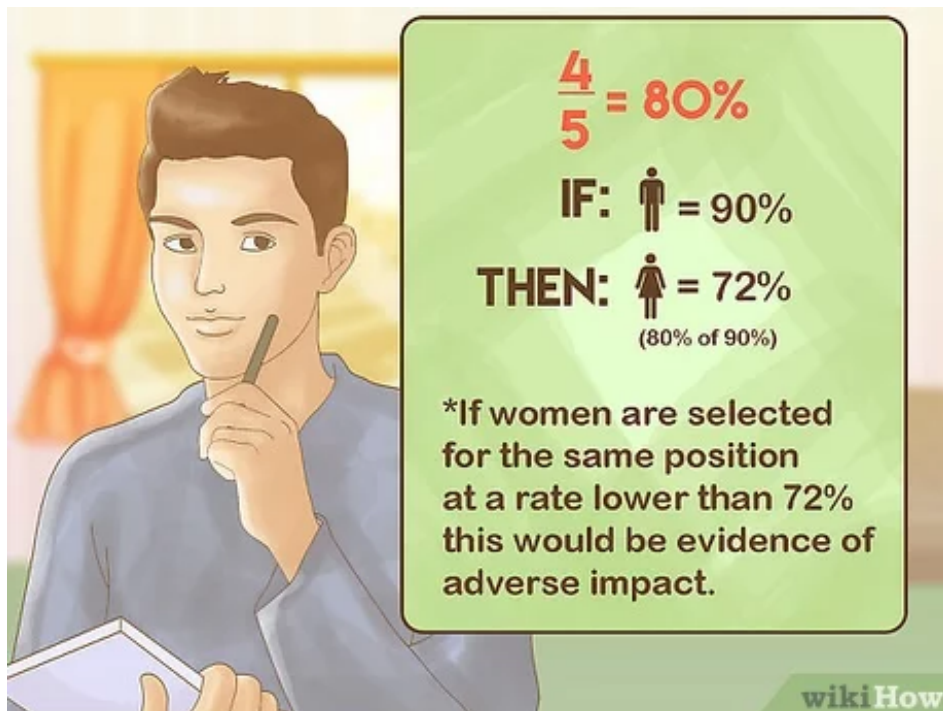
Now, before we begin explaining some of the techniques that have been developed to address the issue, we need to understand that this is a socio-technical topic and therefore new developments are heavily influenced by the prevailing law. A key idea is that practitioners really have only one reason to address these issues from a non-ethical standpoint: being incriminated in a discrimination lawsuit.

US Law has two main criteria for defining discrimination: *disparate treatment and disparate impact*. Disparate treatment is the straightforward idea that any decision which explicitly disadvantages a particular protected group (based on race, gender, etc) is unconstitutional and therefore a grounds for discrimination. Disparate impact is the more nuanced idea that discrimination may sometimes occur even without any explicit reference to the protected group. It refers to the consequences of the decision as being unfair.


The way that disparate impact is contested is an important part of some of the literature in the field of algorithmic de-biasing. While there is no specific numerical formula to define disparate impact, the “4/5ths or 80% rule” is often cited as being the grounds for a case to be accepted in court. The 80% rule states that the number of individuals of the protected group assigned to the advantaged decision (or say, the positive class) must be atleast 80% of the number of people assigned that decision from the group with the highest number.


Examples of Disparate Treatment and Impact	
Disparate Treatment	Disparate Impact
Direct discrimination	Indirect discrimination
Unequal treatment	Unequal consequences or results
Decision rules with a racial /sexual premise	Decision rules with racial / sexual consequences
Intentional discrimination	Unintentional discrimination
Prejudiced actions	Neutral, color-blind actions
Different standards for different groups	Same standards, but different consequences for different groups

Business and Society: Ethics and Stakeholder Management, 7e • Carroll & Buchholtz  
Copyright ©2009.



**$\frac{4}{5} = 80\%$**

**IF:**  = 90%

**THEN:**  = 72%  
(80% of 90%)

**\*If women are selected for the same position at a rate lower than 72% this would be evidence of adverse impact.**

wikiHow

If a candidate can demonstrate that the employer failed to comply with the 4/5ths rule, the employer is then given a chance to justify the disparate impact under the “business necessity clause” where they may claim that a certain level of disparate impact is necessary to meet performance related constraints. Hence, from the point of view of the employer, these two ideas need to be kept in mind when they attempt to debias the classifier.

One problem with the 4/5ths rule is that it is not a convex function of the parameters that the classifier model may be using. Therefore, using this definition and incorporating it in machine learning algorithms that generally demand convex problems is not straightforward. The workaround for this is to use some other function, that closely follows the rule but has the added advantage of being convex.

Another issue with both these definitions of discrimination is that they can sometimes be at odds with one another. An algorithm that does not consider the protected attribute is clear of the disparate treatment charge, but it can get implicated in disparate impact if the outcomes of the model are not balanced. However, if the employer takes steps to remove bias, they may end up explicitly using the protected identity of a candidate and thus getting implicated in disparate treatment.

## Key Idea

“

***A possible solution for this problem may be found if we can somehow include the fairness constraints as part of the classifier training. We can then train the classifier to be fair using the protected identity for debiasing but not use it during actual testing.***

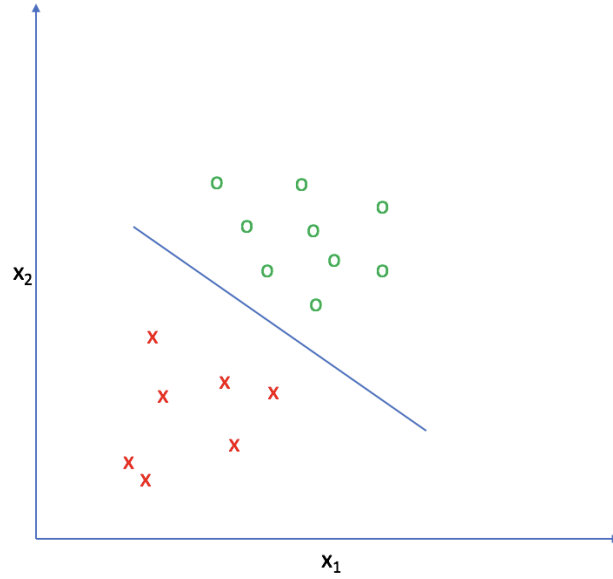
”

Such a solution would circumvent this problem and satisfy both the conditions of fairness according to the law.

## FORMALIZING

The classification problem is then setup with ‘ $\mathbf{x}$ ’ as all the features of the candidate (our knowledge representation from available data) except the protected identity, ‘ $\mathbf{y}$ ’ as the binary outcome (hired = 1 or rejected = 0) and ‘ $\mathbf{z}$ ’ as the protected attribute of the candidate (race/gender/etc.). Now, consider that the classifier uses a decision boundary, i.e. it attempts to find a boundary in the feature space that can separate the two classes ( $\mathbf{y}=1$  and  $\mathbf{y}=0$ )

A view of such a classifier with just two features ( $x_1$  and  $x_2$ ) is in the figure below.



Using this setup, a reasonable measure of fairness can be the correlation between the signed distance from the decision boundary (which models “how far” the point is from being classified as positive/being hired) and the protected attribute  $\mathbf{z}$ . If the signed distance is considerably more for one group over the other, we can say that one group is consistently closer to being classified as being “hireable” over the other. This is related to the notion of disparate impact.

Formally, this is achieved by using the covariance function between  $\mathbf{z}$  and the distance  $\mathbf{d}(\mathbf{x})$  which is also a function of the model parameters  $\boldsymbol{\theta}$ .

$$\begin{aligned} \text{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\approx \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}_i), \end{aligned} \quad (2)$$

Note that, if a decision boundary satisfies the 100% rule, i.e. exactly the same number of positive predictions for both the classes, the covariance will be approximately zero.

As described before, we will use this function to change our training for the classifiers, but our knowledge of this function will make no difference at test time. Infact, to satisfy disparate treatment, we will specifically not use the  $\mathbf{z}$  value at test time.

## SOLUTION STRATEGIES

Now, keeping in mind the general idea described above, we want to incorporate this definition of fairness in the training process. This in turn can be done in two main ways:

### SOLUTION #1: Maximizing accuracy under fairness constraints

This takes the form:

maximize **accuracy**  
such that **fairness is satisfied**

We usually maximize accuracy by minimizing some defined loss function. Hence, this idea reduces to the following formulation where  $L(\theta)$  defines the loss.

$$\begin{aligned} &\text{minimize} && L(\theta) \\ &\text{subject to} && \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\theta}(\mathbf{x}_i) \leq \mathbf{c}, \\ &&& \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\theta}(\mathbf{x}_i) \geq -\mathbf{c}, \end{aligned}$$

This idea is to comply with the 4/5ths rule. The parameter ‘c’ controls the disparate impact and thus can be set accordingly to achieve the desired “p%” disparate impact. Basically, what this says is that as we vary c, we vary the amount of correlation that we allow and hence, the amount of disparate impact we treat as acceptable.

### SOLUTION #2: Maximize fairness under accuracy constraints

As before, this takes the reverse form:

maximize **fairness** OR minimize **covariance**  
such that **accuracy is within control**

Formally, this can be written as below.

$$\begin{aligned} &\text{minimize} && \left| \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\theta}(\mathbf{x}_i) \right| \\ &\text{subject to} && L(\theta) \leq (1 + \gamma) L(\theta^*), \end{aligned}$$

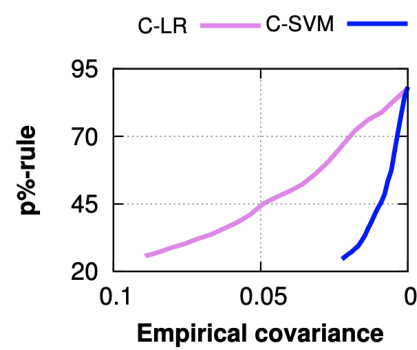
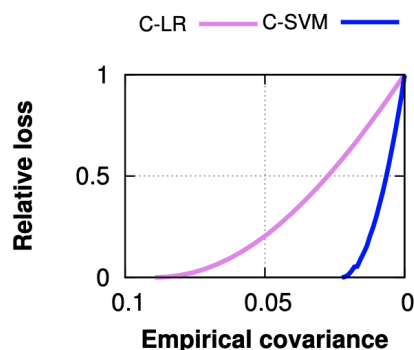
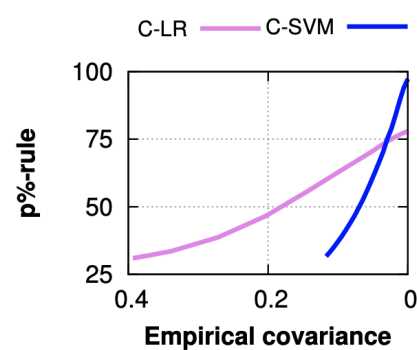
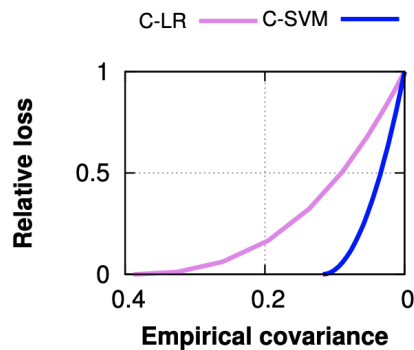
This idea is related to the “business necessity clause”. Here we say that, we want to minimize the level of disparate impact created by the decision while keeping some baseline for the performance. Basically, what this says is that as we vary  $\gamma$ , we are varying the utility of the

classifier or the performance of the whole model itself. Now, since we don't want to go beneath a certain performance level, the minimization itself gives us the best that we can do in terms of disparate impact. This can serve as justification for the employer and hence, still be fair in the eyes of the law.

## RESULTS

Both these solution strategies have the advantage of being convex problems due to our formulation of bias and hence have easy and defined ways of implementation for different  $L(\theta)$ 's. We use two popular classifiers: logistic regression and support vector machines to modify using our additional constraints and do performance analysis for different  $c$ 's and  $\gamma$ 's.

The datasets used for the same are "Adult Income Dataset" (UCI) and the "Bank marketing Dataset" (UCI). The Adult dataset contains 45,222 subjects, each with 14 features and a binary label which says if the subject's income is above/below \$50K. The binary protected attribute here is gender which can be male or female, with females as the protected group. The Bank dataset has 41,188 subjects, each with 20 features and a binary label which says if the subject has subscribed or not to a term deposit. The protected attribute here is age, and it is binarized using the threshold of 25, i.e. ages less than 25 and ages more than 25.



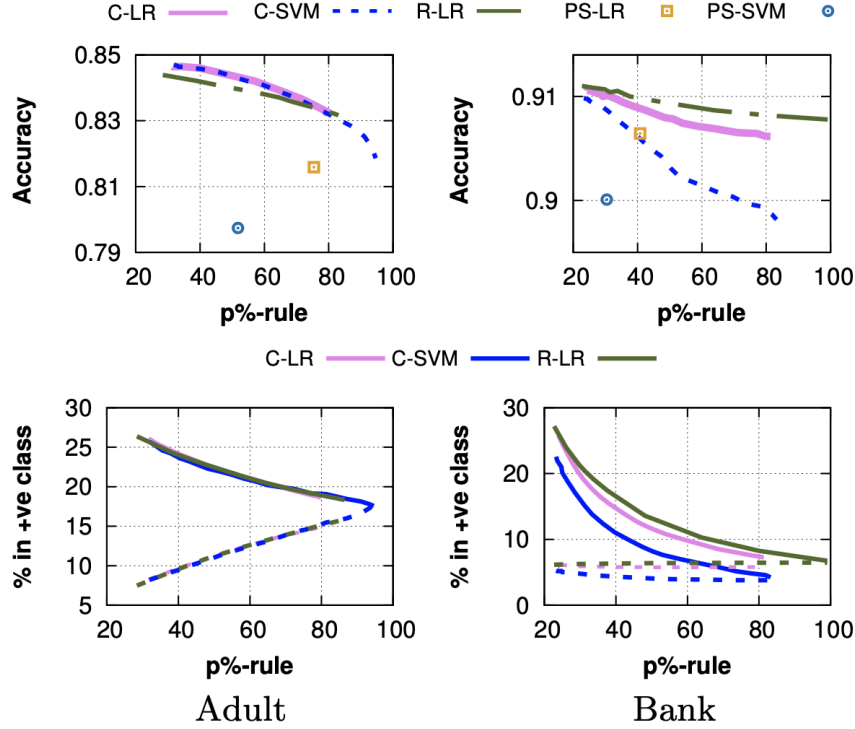
Adult and Bank

Adult and Bank

(a) Loss vs. Cov.

(b) Cov. vs.  $p\%$ -rule

First the results using solution strategy 1 are presented, i.e. maximizing accuracy with fairness constraints. The results for the two datasets are in the given graph. The top is for the Adult database and the bottom is for the Bank. From graph (a), we can see that as we constraint the covariance to be closer to 0, we get increasingly higher loss, or lack of accuracy. From (b), we can see that this change in covariance is reliably related to disparate impact, since as we constraint the covariance closer to 0, we get lesser disparate impact or satisfaction of a higher p% rule.

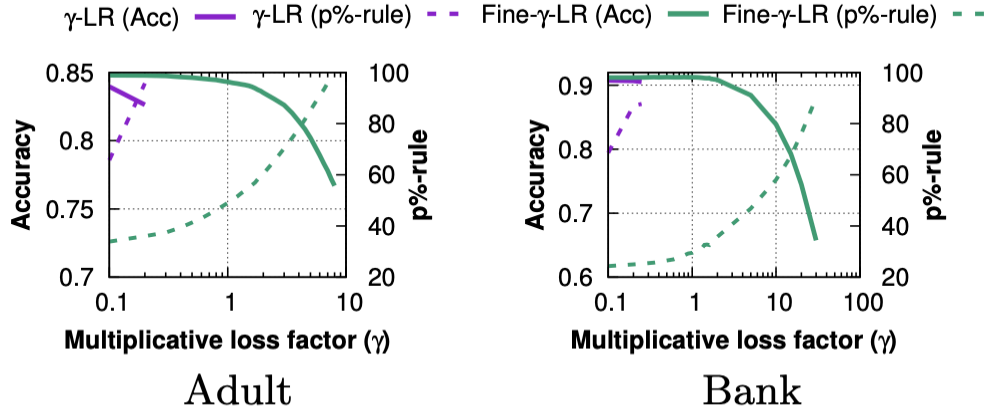


(c) Single binary sensitive attribute

In graph (c) we compare this technique against two other techniques marked as “PS-” and “R-LR” where PS refers to the preferential sampling approach developed in Kamiran and Calders (2010) and R refers to the regularization approach developed by Kamishima et.al. (2011). We can see that the regularization approach is comparable to the proposed method (they are almost theoretically equivalent) and that the proposed approach performs better (higher accuracy) than preferential sampling based debiasing consistently.

The advantage that the proposed method has over regularization (2011) is that unlike it, the current method does not use protected attribute  $z$  during test time hence satisfying both disparate treatment and disparate impact.

The results for strategy 2, minimizing disparate impact for given accuracy at various different  $\gamma$  are given below:



(a) Acc and  $p\%$ -rule

From this graph, we can see that as we allow  $\gamma$  to be larger, we are restricting the accuracy but increasing the p% rule, i.e. decreasing the disparate impact. The key idea of solution strategy 2 is not in this variation though, this merely confirms the expectations. The utility of strategy 2 is in confirming that this is the “best” that we can do if we allow for a maximum of  $\gamma$  amount of drop in accuracy.

## CONCLUSIONS

Here we presented a method to reduce bias in a classification setting and basically find an intuitive, portable and convex solution strategy for balancing fairness and accuracy. An important thing to note is that the utility of this analysis is not limited to the two strategies presented. Using the general idea, this can be extended to any supervised learning setting and made to use different formulations of bias.

Another useful feature is that it considers all aspects of fairness law which is, in truth, the only requirement that a lot of algorithmic decision makers want to fulfil. Hence, unless the law is updated, this work will stand the test of time.

There are a few important issues as well, though. First, this technique uses distance based classifiers but does not talk at all about robustness to adversaries. Second, its formulation of bias is not directly recognizable as a fairness metric. Even though it is generally “related” to the idea of fairness, there is still scope to have a much better formulation. Thirdly, and most importantly, the underlying assumption of this technique may be flawed. This idea is given in the paper: “Fairness and Abstraction in Sociotechnical Systems” discussed below.



## FURTHER READING

On the broader ideas of fairness and debiasing and the issues that most literature, including this one, fails to tackle.

1. [“Fairness and Abstraction in Sociotechnical Systems”, A. Selbst et.al 2019](#)

The key idea of this paper is that any attempt to “fix”, “solve” or “debias” fairness problems by just looking at outcome based fairness measures, and by trying for a generalized solution is in itself flawed. The paper claims that by abstracting away the social context in which an ML system is deployed, researchers miss important information that can lead to fairer outcomes. The paper then goes on to describe 5 traps in which fair-ML literature can fall and how each of them reduce the utility and fairness of the proposed solution. With regards to the paper at hand, it is very clear that the authors have fallen into the most basic trap: the framing trap.

Previous work inspiring and mentioned in this paper.

2. [“Fairness-aware Classifier with Prejudice Remover Regularizer”, Kamishima et.al. 2011](#)

The key idea of this paper (similar to the one hand) is that fairness is simply a constraint that has to be imposed on the learning of the classifier. Unlike this paper, they do not consider the law relating to such issues, but instead discuss the causes of such bias. Also, instead of forming a convex problem, they directly propose a regularization based solution for prediction algorithms that use discriminative probabilistic models. We can see strong theoretical similarities between the two papers in spite of the differences in formulation, when we note that regularization can simply be viewed as a dual of the convex problem. Hence, the present work is some sort of direct extension.

3. [“Classification with No Discrimination by Preferential Sampling”, Kamiran, Calders 2010](#)

The key idea in this approach is that data objects closer to the decision boundary are more likely to be discriminated against since the confidence of the model in their classification is low. The authors therefore use a heuristic to resample the database based on rankings of the distance from the decision boundary. This is a pre-processing approach which, though an interesting innovation at the time, is not well favoured today because of its intrusiveness in the data. Also, the results that were obtained using this approach were easily bettered using the paper at hand.

Other approaches to this domain

4. [Censoring Representations with an Adversary. Edwards and Storkey. 2016](#)

The key idea of the paper is that the data needs to be de-biased by learning a latent embedding of the data specifically tailored to hide all information about the protected attribute. This is thus, essentially a pre-processing technique that works well for a variety of problems. This latent representation is obtained by using the data to simultaneously train an adversary to predict the protected attribute from the latent representation. Its ability to do so is then quantified and added to the combined loss.

5. ["The Variational Fair Autoencoder", Louizos et.al. 2017](#)

The key idea of this interesting paper is to learn a "purged" representation of the data wherein any relation to the sensitive attribute has been removed using the variational autoencoders technique. The paper also suggests a regularization technique to remove any leftover effect of the protected attribute after the pre-processing. This, however, is different from the current work in the idea that it does consider the protected attribute in order to get a representation independent of its effect and thus, does not satisfy disparate treatment requirements.