Visualizing and Interpreting Adversarial Fairness Regularizers

Xiaomeng Hu
University of Toronto
Toronto, ON
xiaomeng.hu@mail.utoronto.ca

Lun Yu Li
University of Toronto
Toronto, ON
lunyu.li@mail.utoronto.ca

Sidharth Gupta
University of Toronto
Toronto, ON
sid.gupta@mail.utoronto.ca

Abstract

A remarkable byproduct of machine learning is how it's pushed the scientific community to define the idea of fairness, in terms of probability and logic. Such definitions are motivated by empirical results, but in the abstract they can be used in every field of science, to make them more accessible, justified, and united. In this paper, we visually interpret two algorithms that make machine learning models follow probabilistic fairness definitions. Our main contribution, though, is visualizing how these fairness definitions translate to differences in the weights, principle components, and latent representations of models. Our results show that visually, these fairness definitions bring models to put less stress on minority groups, which is the desired philosophical outcome. We hope that our work can make probabilistic fairness a more digestible concept to understand, and can encourage scientists in other fields to think about fairness in terms of data, weights, principal components, and latent representations.

1 Introduction

1.1 Fairness and why it's important

Machine learning classification has revolutionized many areas of science – healthcare, economics, and finance to name a few. However, a supervised classification model is almost entirely dependent on it's dataset, which can inherently represent biases that exist in the real world. For example, state-of-the-art deep learning classifiers that predict diagnostic labels from X-ray images show significantly more true positives for males than females [7]. A model that has such a disparity in performance between male and female groups possibly has not learned to highlight meaningful features for the task, which is especially alarming in a healthcare domain. This defines our intuitive idea of fairness: when the same model performs significantly better on one preferred group versus another. Investigating machine learning fairness is an effort in interpretability, because we want to make sure our models do not overfit and fixate on biases in the dataset. And it's important that this research continues, otherwise, the biases that we see in our datasets will propogate to the machine learning models we build, and further deepen the inequality that we see in society.

1.2 The UCI dataset

This paper makes fairness a more digestible concept through visualization, and we'll create visualizations by running experiments on the UCI dataset [3]. The task of this dataset is to predict if the income of a person is more than $50k, given attributes such as sex, race, education, age, previous occupation, etc. 67% of the data in the dataset comes from men, with 30.6% of men labelled with a salary of more than 50,000. On the other hand, 33% of datapoints come from women, with only 16.5% of women labelled with a salary greater than 50,000. It's clear to see that there is a large bias in how this dataset was collected, and as such no model should be trained on it without fairness engineered inside.

1.3 Visualizing Unfairness

We'll begin by showing what this large dataset bias looks like, in the data-space and latent-space. Figure 1 shows the PCA plot of the original dataset, and presents two clear clusters for each gender (male blue, female red). It also shows the PCA plot of the latent space from an unfair MLP model, and this latent shows a large dense male cluster, with a much more varied female cluster. As a result, it is easy for the unfair MLP to predict from the male cluster, since it has a low variance, but it is hard for it to predict from the female cluster since it has a much higher variance.

PCA of the UCI dataset PCA of the latent from an unfair MLP

Figure 1: We present two Principal Component Analysis (PCA) figures, with blue dots representing male points and red dots representing female points. The unfair model is trained using a 3-layer MLP, with a size 113 input layer, size 8 hidden/latent layer, and size 1 output layer.

In Figure 2, we take the size (113, 8) weight matrix mapping inputs to latent in the unfair MLP, and sum elements across the column-axis (resulting in a size 113 vector, the same size as the number of features). The figure shows that the female and male gender features (row 68 and 69, circled in red) have comparatively large weights. In fact, the gender features have a significant impact when compared to the occupation weights on the predictions. The summed male weights are higher than 21% of the summed occupation weights, and the summed female weights are higher than 85% of the summed occupation weights. Thus, the model weighs gender similarly if not more than occupation when predicting income, which does not make any sense. This figure presents how unfairness would greatly influence the predictions of a machine learning model and emphasizes the importance to dampen the weights on protected features with regularization techniques.

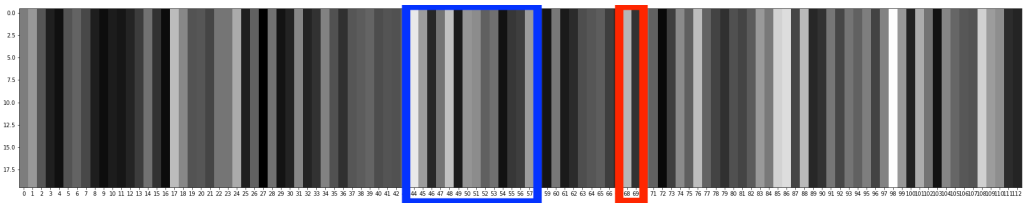


Figure 2: We visualize the weight matrix of the unfair MLP's first layer, summed across the column-axis (resulting in a size 113 vector, same size as the number of features). We increase the height of each bin for better visualization. Weights associated with gender are in the red box, and weights associated with occupation are in the blue box.

2 Background material

Now that we know what unfairness looks like, we can introduce some probabilistic definitions of fairness. There are many metrics we can use for evaluating fairness of a model. Some of the most well-known ones are Disparate Impact, Demographic Parity, Equality of Odds and Equality of Opportunity. Let P(Y = $\hat{y}$) be a probability mass function that represents our model making a prediction $\hat{y} \in \{0, 1\}$, and let Z $\in \{0, 1\}$ be a binary sensitive feature.

2.0.1 Disparate Impact (DI)

Disparate Impact (DI) is a fairness metric given by comparing the number of positive output Y for the unprivileged group and the privileged group. I.e. $\frac{P(Y=1|D=unprivileged)}{P(Y=1|D=privileged)}$.

2.0.2 Demographic Parity and Demographic Parity Distance (DP)

We say a model has Demographic Parity if its predictions are independent of the sensitive feature. I.e. $\forall y \in \{0, 1\}, P(Y = \hat{y}|Z = 0) = P(Y = \hat{y}|Z = 1)$. Demographic Parity Distance (DP) is a fairness metric given by DP = $||P(Y = 1|Z = 0) - P(Y = 1|Z = 1)||$

2.0.3 Equality of Odds

We say a model has Equality of Odds if its predictions are independent of the sensitive feature, given any target. Concretely: $\forall y \in \{0, 1\}, P(Y = \hat{y}|Y = y) = P(Y = \hat{y}|Z = z, Y = y)$

2.0.4 Equality of Opportunity

Equality of Opportunity is very similar to Equality of Odds, except that instead of any target, Equality of Opportunity enforces the predictions to be independent of the sensitive feature given a particular, preferred target. Concretely: $\exists y \in \{0, 1\}, P(Y = \hat{y}|Y = y) = P(Y = \hat{y}|Z = z, Y = y)$

2.0.5 Positive Rate and Negative Rate

Positive rate calculates the ratio of number of prediction $\hat{y} \geq 0.5$ over total number of predictions. Negative rate calculates the ratio of number of prediction $\hat{y} < 0.5$ over total number of predictions. Additionally, the sum of positive rate and negative rate should be 1.

2.0.6 False Positive and False Negative

False positive is the total number of incorrect prediction $\hat{y} \geq 0.5$ given for a true $y = 0$. False negative is the total number of incorrect prediction $\hat{y} < 0.5$ given for a true $y = 1$. False positive and false negative consider all cases for incorrect predictions.

2.1 Adversarial Learning Algorithms

We will visually interpret two adversarial algorithms: Adversarial Debiasing and Adversarial Representation Learning (LAFTR). Let X be the input variable, Y be the output variable, and Z be the sensitive features.

2.1.1 Adversarial Debiasing

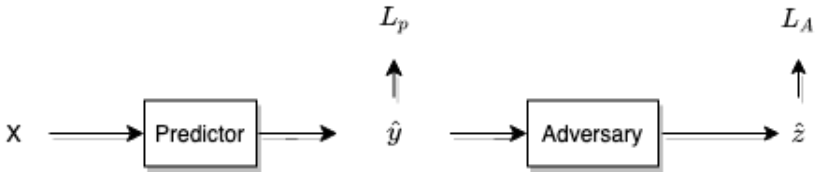


Figure 3: The architecture of Adversarial Debiasing model[8]

Adversarial Debiasing was proposed by Zhang et al. in 2018.[8] The method has two models: the predictor and the adversary. The predictor predicts Y given X, while the adversary tries to predict Z bases on the output of the predictor. During training, the predictor not only targets to increase its prediction accuracy, but also tries to increase the adversary's loss so that it cannot predict the sensitive features well. [8]

Figure 3 gives a representation of the overall structure of the debiaing network, where L_P is the loss of the predictor and L_A is the loss of the adversary. The gradient by which the weights of the predictor, W, are updated is:

$$\Delta_W L_P - proj_{\Delta_W L_A} \Delta_W L_P - \alpha \Delta_W L_A$$

where α is a hyperparameter. The first term $\Delta_W L_P$ tries to decrease the predictor loss, while the following terms $proj_{\Delta_W L_A} \Delta_W L_P - \alpha \Delta_W L_A$ attempts to increase the adversary loss.[8]

Zhang et al. also proved that under certain conditions, if the predictor and adversary converge, then the converged model will satisfy the fairness constraints, such as demographic parity and equality of odds.[8]

2.1.2 Adversarial Representation Learning (LAFTR)

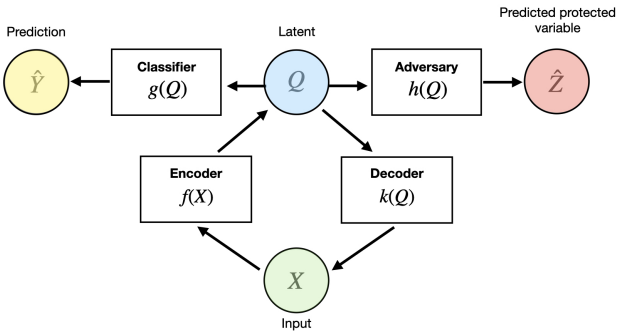


Figure 4: The architecture and components of Adversarial Representation Learning (LAFTR).

Madras et al. proposed the above LAFTR architecture in 2018.[6] LAFTR is used for training a powerful encoder by maximizing the loss for the adversary while minimizing the loss for the classifier and the decoder. Intuitively, we want the adversary to make as bad predictions of the protected

variable as possible. We use different loss functions for the different criteria we want to represent. Let D_i represent the dataset for when the protected variable $z = i$ for z $\in \{0, 1\}$.

For Demographic Parity, we compute the average difference between the adversary getting the protected variable and the actual protected variable z. This loss function is defined to be

$$L_{Adv}^{DP}(h) = \sum_{i \in \{0,1\}} \frac{1}{|D_i|} \sum_{(x,z) \in D_i} |h(f(x)) - z|$$

For Equality of Odds, let D_i represent the dataset for when the protected variable $z = i$ and label $y = j$. This is similar to the demographic parity loss function, but we are specifically considering the case where we know both the protected variable and the label. The loss function is defined to be

$$L_{Adv}^{DP}(h) = \sum_{i \in \{0,1\}^2} \frac{1}{|D_i^j|} \sum_{(x,z) \in D_i^j} |h(f(x)) - z|$$

For Equality of Opportunity, let D_i represent the dataset for when the protected variable $z = i$ and label $y = j$. This is similar to the equality of odds, but we are fixing a specific variable j for the label. The loss function is defined to be

$$L_{Adv}^{DP}(h) = \sum_{i \in \{0,1\}, j=1} \frac{1}{|D_i^j|} \sum_{(x,z) \in D_i^j} |h(f(x)) - z|$$

3 Related work

The fair spectral clustering has been discussed in the previous work. Chierichetti et al. [2] introduced fairlets as an appropriate notion of fairness and showed that any fair clustering problem can be solved by finding good fairlets and solve it with existing clustering algorithms. They also provided approximation algorithms to find good fairlets because the problem is NP-hard.

Kleindessner et al.[5] presented an algorithm that incorporates fairness constraints into the spectral clustering framework and help find fairer clusterings.

Adel et al. [1] proposed the one-network adversarial framework. Instead of building a model from scratch, they add a hidden layer and a classifier to an existing classifier and significantly improved the fairness of the model. Their work is very intriguing and it would be interesting to interpret their model to see how fairness improves with training. We will not do it in this report given the limited space.

In the report, we will interpret the LAFTR and Adversarial Debiaing methods and try to visualize the fairness with PCA to see how the fairness coefficient can affect the prediction results in terms of probabilistic fairness and accuracy.

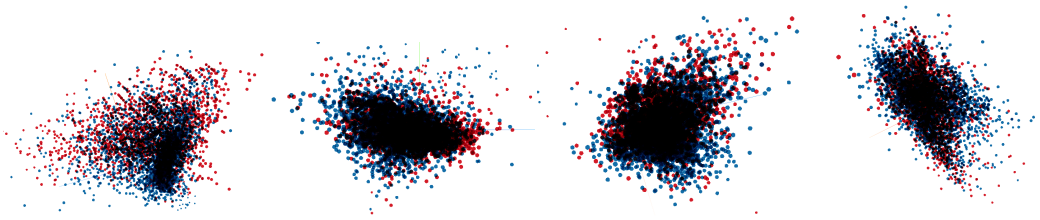
4 Experiments

At this point, we've seen what unfairness looks like, and have talked about probabilistic definitions of fairness. We'll now conduct experiments that apply these probabilistic definitions to unfair models, and show how they visually change the weights, latent representations, and principal components.

4.1 Adversarial Representation Learning (LAFTR)

We'll start by defining an experiment on the LAFTR model. We used a fixed architecture for LAFTR, which is an MLP with input size 113, latent size 8, and output size 1. As mentioned 2.1.2, there will be an adversary who tries to predict the protected variable from the latent representation. For brevity, we'll include experiments using the Adversarial Demographic Parity loss function, as we'd expect to find similar results given the other fairness loss functions. The role of the adversary's loss during training is scaled by a fairness coefficient, which we'll denote as α.

In Figure 5, we visualize the PCA plot of trained LAFTR models with varying α.



$\alpha = 0$ $\alpha = 0.5$ $\alpha = 1.0$ $\alpha = 1.5$

Figure 5: For each α, an MLP LAFTR model is trained, with a PCA plot created on the latent data representations

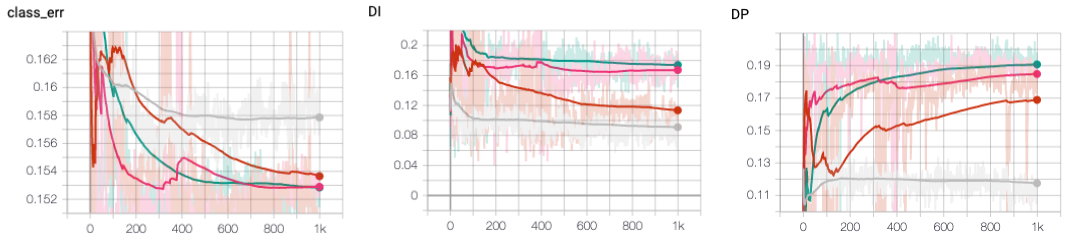
When $\alpha = 0$, we can see that the MLE creates a strong male cluster with a lower variance, which is evidently easier to classify from. The female cluster, however, is much more spread with higher variance, making it harder to classify from. We see this problem mitigated when $\alpha = 0.5$ and $\alpha = 1.0$, as the variances seem to equal out, and there's no prominent gender cluster. When $\alpha = 1.5$, both the male and female clusters become more varied, which is okay since they look equally varied, but can hint at a reduction in classification accuracy, as this much variance makes it harder to classify from.

Next in Table 1, we display the proportion of summed weights at varying fairness coefficients. When $\alpha = 0$, there is a large proportion of occupation weight sums that are less

	% summed occupation weights less than male weights	% summed occupation weights less than female weights
Fairness coefficient 0	0.21	0.85
Fairness coefficient 0.5	0.21	0.00
Fairness coefficient 1.0	0.07	0.00
Fairness coefficient 1.5	0.07	0.00

Table 1: For the model at varying α, we save the size $(113, 8)$ weight matrix going from input to latent, and sum across the column-axis (getting a size 113 vector, equal to the number of input features). Then, we compare the proportion of occupation-related weight sums (weights 43 through 57) that are less than gender weight sums.

than female weight sums, which gives evidence that the model favours the "female gender" as a feature more than occupation. Of course, this doesn't make sense for the task of income prediction. Setting $\alpha = 0.5$ removes this problem, and female weight sums are less than all occupation weight sums. When $\alpha = 1.0$ or $\alpha = 1.5$, the dependence on the male weight sums decreases.



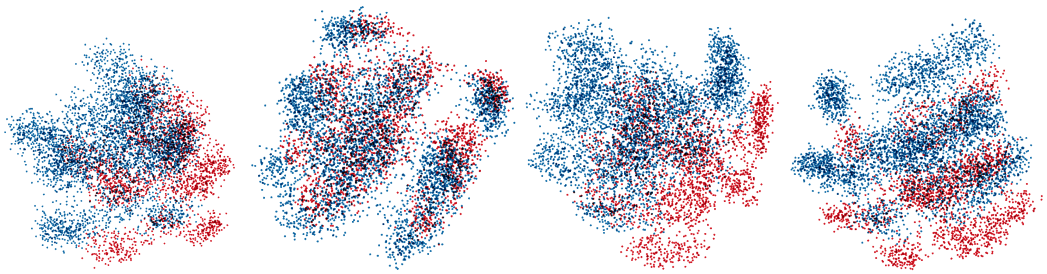
Classification error DI Values DP Values

Figure 6: We present training plots that articulate the classification error, DI, and DP values of the LAFTR model, at varying α. Gray: $\alpha = 1.5$; Red: $\alpha = 1.0$; Pink: $\alpha = 0.5$; Green: $\alpha = 0$

In this experiment, we see that as α increases, DI and DP (mentioned in 2.0.1, 2.0.2) decreases, and classification error increases. This is what we would expect, as the model does a bit worse at classifying the majority group (males) on the test set, but would do much better at classifying the minority group (females).

4.2 Adversarial Debiasing

We will perform the same set of experiments, except using the Adversarial Debiasing model [8]. These plots are a bit trickier to interpret, and that is because Adverserial debiasing is not a data



$\alpha = 0$ $\alpha = 0.5$ $\alpha = 1.0$ $\alpha = 1.5$

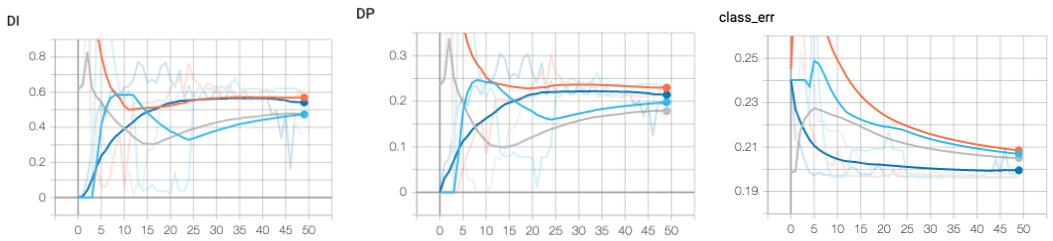
Figure 7: For each α, an MLP model is trained with adversarial debiasing. A PCA plot is created on the latent data representations

representation-learning algorithm like LAFTR. Still, in Figure 7 we can see that when $\alpha = 0$, there are more dense male clusters than female, and when $\alpha = 1.5$, both male and female clusters look quite similar (they both follow a similar "horizontal row" like pattern). This table shows the same

	% summed occupation weights less than summed gender weights
Fairness coefficient 0	0.66
Fairness coefficient 0.5	0.55
Fairness coefficient 1.0	0.21
Fairness coefficient 1.5	0.00

Table 2: Analogous to Table 1. Here, we save a size (18, 200) weight matrix going from input to latent, and sum across the column-axis to get a size 18 vector (which is our number of input features). Both male and female are represented through one feature dimension, as gender.

behavior as in Table 1, except more prominently. Again, this can be attributed to the fact that Adversarial Debiasing is model-driven, and thus places more emphasis on regularizing the weights when training to be fair.



Classification error DI Values DP Values

Figure 8: We present training plots that articulate the classification error, DI, and DP values of the LAFTR model, at varying α. Orange: $\alpha = 0$, Blue: $\alpha = 0.5$, Gray: $\alpha = 1.0$, Cyan: $\alpha = 1.5$

Again, we see the same behavior as in Figure 6. As α increases, DI and DP decrease, while classification error increases. These results look similar, because DI, DP, and classification error are metrics that we compute after training and testing is done. So it shows that both LAFTR and Adversarial Debiasing are two different ways of accomplishing the same goal.

5 Discussion

5.1 The impact of fairness

In our experiments, we observe three results as the fairness coefficient α increases: first, intense weights on protected variables such as gender decreases to something that is reasonable and makes intuitive sense. Second, latent representations that have dense majority clusters and varied minority clusters even out to clusters with similar variance. And third, as α increases, we see a reduction in DI and DP, but an increase in classification error (likely due to more majority samples being classified incorrectly, even though more minority samples will be classified more correctly).

5.2 When to use which

There are usually two parts in a prediction task – obtaining data and making predictions based on the data. In reality, these two parts are often carried out by different parties. After the data owner collects the data, it can be sold to prediction vendor to make predictions based on the given data.[4]

LAFTR provides a method of learning adversarially fair representation, which the data owner can use to present the data more fairly before selling it to the prediction vendors. Meanwhile, the accuracy of the final predictions will not be affected.

After the prediction vendors have acquire the data from data owner, they can use adversarial debiasing to make more fair predictions. Note that the usability of adversarial debiasing is very extensive. As long as the original model is trained with a gradient-based method, adversarial debiasing can be used to make predictions more fair. In addition, in many cases, even if the original model is very complex, you can still use a relatively simple adversary to improve the fairness of the model, which greatly lowers the threshold for debiasing. [8]

In conclusion, LAFTR is Data-representation driven and Adversarial Debiasing is Model driven. When used in appropriate situations, they can both improve the fairness of predictions without greatly reducing accuracy.

6 Conclusion

In this work, we perform an interpretability study on probabilistic fairness. Through the lens of the highly biased UCI dataset, we visually show what unfairness looks like using PCA and unfair MLP models. Specifically, we visualize the latent space of these unfair MLP models using PCA, and also count the proportion of summed gender weights that are greater than summed occupation weights. We then use adversarial machine learning models to augment these unfair MLPs so that they follow probabilistic fairness definitions developed in the scientific literature. Each of these models depends on a fairness coefficient, which we denote as α, that controls the degree of fairness regularized in the model. We scale α to values $0, 0.5, 1.0, 1.5$, and visually show how the latent MLP PCA, weights, and DI, DP metrics change accordingly. We find that latent representations between minority and majority groups form clusters with equal variances, as opposed to a majority cluster being very dense. In addition, we see that the summed weight values of the protected variable becomes less than the summed occupation weights, showing that they're deemed less important, as desired. And finally, we show that as α increases, DI and DP decreases, but classification error increases (likely because more majority samples are misclassified, even though more minority samples are correctly classified). Ultimately, we show the effects of scaling α numerically and visually. We hope that our interpretation work makes these probabilistic definitions of fairness more digestible for other areas of science to use.

References

[1] Tameem Adel et al. "One-Network Adversarial Fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 2412–2420. DOI: 10.1609/aaai.v33i01.33012412. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4085.

[2] Flavio Chierichetti et al. *Fair Clustering Through Fairlets*. 2018. arXiv: 1802.05733 [cs.LG].

[3] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[4] Cynthia Dwork et al. *Fairness Through Awareness*. 2011. arXiv: 1104.3913 [cs.CC].

[5] Matthäus Kleindessner et al. *Guarantees for Spectral Clustering with Fairness Constraints*. 2019. arXiv: 1901.08668 [stat.ML].

[6] David Madras et al. *Learning Adversarially Fair and Transferable Representations*. 2018. arXiv: 1802.06309 [cs.LG].

[7] Laleh Seyyed-Kalantari et al. *CheXclusion: Fairness gaps in deep chest X-ray classifiers*. 2020. arXiv: 2003.00827 [cs.CV].

[8] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. *Mitigating Unwanted Biases with Adversarial Learning*. 2018. arXiv: 1801.07593 [cs.LG].