February 05, 2025, Lecture Notes: Impossibility Theorems; Fairness Interventions

Jirobe, Nandini Satyanarayan[1], Goyanka, Parikha[2]
University of Illinois Chicago, Chicago, IL

1 Introduction

In the previous class, we discussed different categorizations of fairness, focusing on three key groups: independence, sufficiency, and separation. Independence does not consider the target variable or true label. It simply requires that the sensitive attribute be independent of predictions. Sufficiency states that predictions should be independent of sensitive attributes but only when conditioned on the true label. Separation requires that the sensitive attribute be independent of the true label when conditioned on predictions.

We then explored impossibility theorems, which assume that bias exists in the data. This means that if the target variable Y is dependent on the sensitive attribute (indicating bias), it is mathematically impossible to satisfy more than one of the three fairness definitions simultaneously. In other words, we must choose between independence, sufficiency, or separation, as it is impossible to achieve all three at once. We concluded by discussing one of the few impossibility theorems: that independence and separation cannot be achieved together. In this lecture, we are going to continue learning about other impossibility theorems, ways of measuring unfairness, properties of a database, and interventions to reduce bias.

2 Independence and Sufficiency are Mutually Impossible

One of the second impossibility theorems is that Independence and Sufficiency are mutually exclusive. In other words, you cannot achieve both of these categories of fairness at the same time. To prove this, let us try going over a simple contrapositive proof.

In this proof, let S be the sensitive attribute, F be predictions, and Y be the target variable. Sufficiency is the class of definitions that require sensitive attributes to be independent of predictions, but only if you condition the target variable. In other words:

$$S \perp F \mid Y$$

Independence is the class of definitions that requires sensitive attributes to be independent of predictions. In other words:

$$S \perp F$$

Assume that S and Y are not independent. Then, sufficiency and independence cannot both hold.

[1] njiro2@uic.edu
[2] pgoya45@uic.edu

Based on the definition of Sufficiency, if you condition on Y, S and F are going to be independent. That means the path S-Y-F (look at Figure 1) would exist.

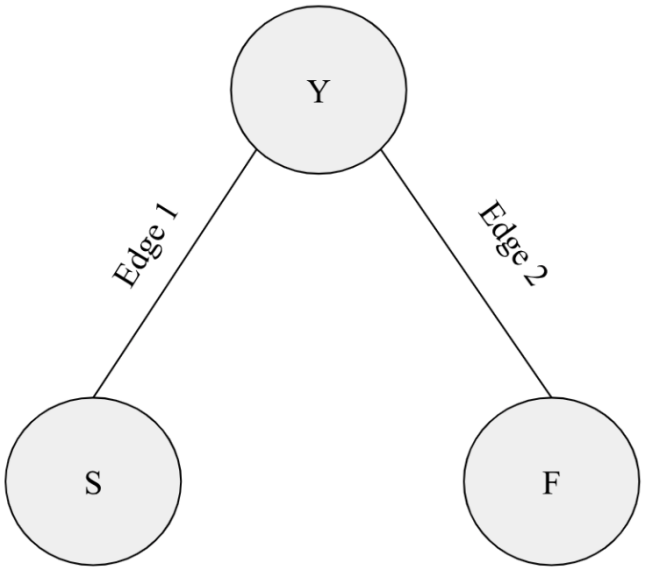

Figura 1: The S-Y-F path

However, if this path exists, Swould not hold true. This is because, according to this path, S is dependent on Y, and Y is dependent on F. If node Y is removed, then that would mean that S would be dependent on F. Hence, the path S-Y-F does not exist.

To make S or F independent, deleting either Edge 1 or Edge 2 would be necessary. If we deleted Edge 1, that would mean S. This is not possible because Bias Assumption, which states that S and Y are dependent on one another. Hence, this edge already exists. Alternatively, if we deleted Edge 2, that would mean Y. However, this suggests that predictions have no effect on the target variable – which is meaningless because there is a correlation between the predictions and target variables.

If both Edge 1 and Edge 2 exist, then the path S-Y-F also exists. If this path exists, then this shows that S and F cannot be independent of each other, hence proving that Independence and Sufficiency cannot happen at the same time.

3 Separation and Sufficiency are Mutually Impossible

Sufficiency is the class of definitions that require sensitive attributes to be independent of predictions, but only if you condition on the target variable. In other words:

$$S \perp F \mid Y$$

Separation is the class of definitions that requires sensitive attributes to be independent of the target variable, but only if you condition on the predictions variable. In other words:

$$S \perp Y \mid F$$

If both conditions can be achieved, then that would mean that S is independent from the joint distribution of Y and F:

$$S \perp F \mid Y \quad \text{and} \quad S \perp Y \mid F \quad \Rightarrow \quad S \perp (Y, F)$$

However, this would mean that SF it is false because it violates the Bias assumption. This proves that it is not possible to achieve both Separation and Sufficiency.

4 Summary of Impossibility Theorem

The impossibility theorem in fairness states that you cannot simultaneously satisfy all three fairness criteria (Independence, Separation, and Sufficiency).

With that said, let's revisit the Case Study: Recidivism Scores and COMPAS discussed in earlier lectures. ProPublica argued that the scores violate demographic parity (which falls under the Independence category) and are, therefore, unfair. Northpointe claimed that, among those who received the same risk score, the likelihood of reoffending was similar across racial groups—meaning the system was fair under Sufficiency. Supreme Court said that out of the ones that are predicted as positive, the chance of really being positive was the same across the two groups (which means it is fair from the Separation perspective). How is this possible, given that we know all three fairness criteria cannot be achieved simultaneously? This is because Northpointe did not claim perfect fairness across all definitions – only that it was almost equal across all of them, which is possible. However, Northpointe did a better job at reducing the harm to marginalized groups by focusing on improving demographic parity.

Thus, although achieving perfect fairness in all three categories is impossible, we should still aim to minimize unfairness and approximate fairness as much as possible. Impossibility theorems only prove that exact equality among conflicting fairness definitions is unattainable. However, we can still strive for near-equality where trade-offs are carefully considered, ensuring that no group faces excessive disadvantage.

5 Measuring Unfairness

The unfairness measurement of a model or algorithm is a critical and important step to ensure the best practices and overall results. There are mainly two approaches to measure unfairness:

5.1 **Causal Inference Approach:** This method evaluates fairness by considering what the out come would have been if a sensitive attribute had been different. It ensures that the model's decisions remain consistent across different demographic groups by analyzing potential outcomes in hypothetical scenarios.

5.2 **Fairness Metrics (beyond traditional measures):** This metric requires that a model's prediction be independent of sensitive attributes, conditioned on the actual outcome.

There exist a number of such metrics that can be used to measure unfairness. They are mentioned and discussed below:
(Let y' be the predicted outcome, and S is the sensitive attribute, such that S: 1 ; if belongs to protected group and otherwise 0)

5.2.1 Additive and Multiplicative Metrics:
To quantify disparity between different demographic groups, primary approaches are:

- Additive Metrics (Subtraction based): It assesses the absolute difference in outcomes between groups which is,

$$F = P_1 - P_2 \tag{1}$$

This is not enough for the right context and fair metrics. Hence it should be normalized. Therefore, the normalized form that should be used in practice is:

$$F = \left| \frac{P_1 - P_2}{P_1 + P_2} \right| \tag{2}$$

- Multiplicative Metrics (Ratio/ division based):
  It evaluates the relative difference by computing ratios of outcomes between groups. To improvise it, for practice purpose:

$$F = \frac{\min(P_1, P_2)}{\max(P_1, P_2)} \tag{3}$$

  (Here P1 and P2 are the performances of an Algorithm A for groups g1 and g2 respectively.)

5.2.2 Statistical Parity difference (SPD):
It measures the difference in the positive outcomes of a protected group with the overall population to which the group belongs.

$$\text{SPD} = P(Y' = 1 \mid S = 1) - P(Y' = 1 \mid S = 0) \tag{4}$$

- If SPD is 0 or tends to 0 => equal treatment across the group
- SPD: Positive and Negative => favoritism and bias respectively against the protected group.
- It is generally used for checking demographic parity in settings like hiring, loan approvals and school admissions, scenarios where equal selection rates across groups are desirable.
- Limitation: It ignores the fact that applicants at different groups qualify at different rates.

5.2.3 Disparate Impact (DI):
It evaluates the ratio of favorable outcomes between the protected and unprotected group.

$$DI = \frac{P(Y' = 1 \mid S = 1)}{P(Y' = 1 \mid S = 0)} \tag{5}$$

- If DI is 1 or tends to 1 => parity between groups, if value is less than 1 => potential bias against protected group.
- It is generally used in legal contexts such as equal employment opportunity law, where a ratio of less than 0.8 is considered discriminatory.
- Limitation: It accounts only the ratio of selections, and does not account for qualifications.

5.2.4 Equalized Odds difference (EOD):
It examines the difference in True Positive rates (TPR) and False Positive rates (FPR) between groups.
True Positive rate difference:

$$TPR = P(Y' = 1/Y = 1, A = 1) - P(Y' = 1/Y = 1, A = 0) \tag{6}$$

False Positive rate difference:

$$FPR = P(\ Y' = 1 Y = 0, A = 1)\ \ P(\ Y' = 1 Y = 0, A = 0\ ) \tag{7}$$

- If EOD is 0 or tends to 0 => model is consistent across groups. Deviation suggest potential bias.
- Generally used in predictions related to classification like whether the patient has disease or not. It is used to ensure fair predictive performance across groups in high stakes applications such as healthcare, fraud detection, and criminal justice.
- Limitation: It does not consider the base rates (differences in actual outcome prevalence between groups).

5.2.5 Average Odds difference (AOD):

It provides an aggregate measure by averaging the differences in TPR and FPR between groups.

$$AOD = 1/2 * (TPR + FPR) \tag{8}$$

- AOD =0 => equal predictive performance across the groups.
- Generally useful for binary classification systems where we want a single value summarizing how much fairness exists across different decision outcomes.
- Limitation: Similar to EOD, it does not account for base rates.

6 Analyzing Difference vs. Ratio-Based Metrics:

Difference-Based (Additive) Metrics (SPD, EOD, AOD)

- Best for understanding absolute gaps in selection rates.
- Easier to interpret but can be misleading when base rates differ.

Ratio-Based (Multiplicative) Metrics (DI)

- Best for relative comparisons and legal compliance.
- More stable in contexts where base rates are very different.

Conclusion of the above discussion is that every metric serves a specific purpose and choosing the right metrics depends totally upon the scenario, for example use SPD for fairness check in loan approval, while DI for legal discrimination tests and EOD or AOD for healthcare and criminal justice.

7 Properties of a database

7.1 Data Provenance & Lineage – Knowing the source of data, how it has been collected and processed helps spot potential biases early.. Hence to avoid any such possibility.

7.2 Feature Selection & Representation – Features should be relevant to the task and not correlated with sensitive attributes like race or gender. Even if a model doesn't use race directly, it can still learn bias through proxy variables (such as., postal or zip codes).

In simple terms, a good dataset meets three key fairness rules:

- Enough Samples from All Groups: to prevent bias toward the majority. Example: A hiring dataset with **90% men, 10% women** will favor men.

- Correlation Between Features and Output is Maximized:. Example Income & credit score should decide loans, not random details.
- Correlation Between Features and Sensitive Attributes is Minimized example, Zip codes linked to race could lead to hidden discrimination.

In short, a fair dataset is balanced, relevant, and bias-free.

8 Intervention to reduce/get rid of biases

The simple ways to reduce biases in an algorithm and ML model are as simple as we can think of. They are Human Oversight & Actionable Solutions.

8.1 Human-in-the-Loop Systems – Keeping humans in the decision-making loop helps monitor, catch, and fix bias as it happens. Interactive tools let users adjust model behaviors, making model more transparent & trustworthy.

8.2 Algorithmic Recourse – Giving individuals clear steps to change their inputs (e.g., improving credit score for a loan approval) empowers users and improves fairness.

But exactly how are we implementing it in practice? So, to break it more, we have 3 possible ways:

- **Preprocessing Interventions:** Modifying the training data to remove biases before feeding it into the model. It is considered as the least effective method, hence better to avoid.
- **In-Processing Interventions:** Incorporating fairness constraints or modifying the learning algorithm to ensure fair outcomes during model training (fairness wrapper). This is the hardest method and comes up with practical challenges.
- **Post-Processing Interventions:** Adjusting the model's predictions to achieve fairness after training is complete. It is considered as the best practice in Industry, as well as easy to deploy.

Although none of them are perfect, therefore it's important to understand the scenario and realize the importance of all and then try the best applicable one.