[02/17/2025], Lecture Note: Fair ML: Price of Fairness; Pareto-Optimality in Fairness; Causality: a double-edged sword

[Aditya Acharya, Sanmitha Shetty]

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025

Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

Recap

In the paper *A Reductions Approach to Fair Classification* by Agarwal et al.[1], the authors propose a fairness wrapper to improve fairness in machine learning models. This wrapper can be applied to any black-box training algorithm and iteratively adjusts the model to meet fairness constraints, as long as the algorithm supports weighted samples. The paper introduces a unified fairness template for different fairness notions, such as demographic parity. For each group j, the performance measure is defined as:

$$\mu_j(\theta) = E[g(X, Y, S, h_\theta)|\epsilon(X, f, S)]$$

For demographic parity, the fairness constraint is expressed as:

$$|\mu_j(\theta) - \mu_j^*(\theta)| \leq \tau$$

The problem is then reformulated as a constrained optimization.

$$\min \text{loss}(h_\theta(x), y) \quad \text{subject to} \quad M_\mu(\theta) \leq \tau$$

1 Solving the Saddle Point Problem in Fair Classification

In the paper *A Reductions Approach to Fair Classification* by Agarwal et al.[1], the fairness-constrained optimization problem is formulated as a constrained minimization problem. Mathematically, this problem is expressed as

$$\min_\theta \text{loss}(h_\theta(x), y) \quad \text{subject to} \quad M_\mu(\theta) \leq \tau \tag{1}$$

where $M_\mu(\theta)$ represents the fairness constraints, and τ is the permissible violation threshold. The challenge in solving this problem arises due to the existence of a **saddle point** in the optimization landscape when using standard gradient descent, leading to non-convergent behavior. To overcome this, the constrained optimization problem is rewritten in its **Lagrangian form**:

$$L(\theta, \lambda) = \text{loss}(h_\theta(x), y) - \lambda(M_\mu(\theta) - \tau) \tag{2}$$

where λ is the Lagrange multiplier that enforces the fairness constraint. The optimization then involves simultaneously updating both θ (model parameters) and λ. However, due to the **saddle point nature** of the problem, naive gradient descent may lead to oscillations or divergence. To resolve this, the authors introduce **reweighting via cost-sensitive learning**, transforming the problem into a **weighted classification** framework that iteratively adjusts sample weights to achieve fairness.

In this approach, each training sample (X_i, Y_i) is assigned a weight W_i, which modifies the loss function to

$$\sum_i W_i \cdot \ell(h_\theta(X_i), Y_i) \tag{3}$$

where the weight W_i depends on the **fairness constraint violations**.

To further stabilize the optimization, the **Exponentiated Gradient (EG) method** is used to update λ. The optimization proceeds iteratively by solving a series of cost-sensitive classification problems with updated weights until the fairness constraint violation is below a predefined threshold:

$$\max_k |M_\mu(\theta) - \tau| \leq \epsilon. \tag{4}$$

Thus, by reweighting the loss function based on fairness violations and using exponentiated gradient updates, the saddle point issue is effectively mitigated. This method ensures convergence to a classifier that balances **accuracy and fairness**, and it can be applied to **any black-box classifier** that supports weighted samples.

1.1 Challenges of the Solution

While effective, this approach is not the most efficient as it requires solving multiple cost-sensitive classification problems, making it computationally repetitive. Additionally, the number of iterations required for convergence is theoretically bounded by $O(n^2)$; however, as noted in the paper [1], in practice, it typically converges in a small number of iterations (around 10), making it more practical than its worst-case bound suggests.

2 Price of fairness or Cost of Fairness

The Price of Fairness (PoF) quantifies the trade-off between fairness and performance in decision-making models. It represents the difference in the optimum value between an unfair solution and

a fair solution. In many real-world scenarios, improving fairness can lead to a decrease in overall model performance or accuracy.

Mathematically, the PoF is defined as the difference between the loss achieved by an unfair approach and the loss achieved by a fair approach. In an unfair approach (Approach A), the model minimizes loss without considering fairness constraints, leading to the best possible performance. In a fair approach (Approach B), the model minimizes loss while ensuring fairness, meaning it does not disproportionately favor any group. If we denote the loss of the unfair approach as L_a and the loss of the fair approach as L_b, then the Price of Fairness is given by the equation:

$$PoF = L_a - L_b \tag{5}$$

This metric helps quantify how much accuracy or efficiency is sacrificed to achieve fairness.

A practical example of this trade-off can be observed in Magnetic Resonance Imaging (MRI) models. In this scenario, a model may slightly reduce its coverage to ensure fair treatment across different demographic groups. By doing so, the model prioritizes fairness in medical decision-making, even if it comes at the cost of some precision.

3 Pareto Optimality of Fairness:

One way to assess fairness in decision-making is through Pareto optimality, which seeks a balance between competing objectives: fairness and performance. A model that is Pareto optimal ensures that no further improvement in fairness can be made without compromising performance and vice versa. Pareto Optimality is a set of non-dominated solutions. Pareto-optimal solutions are those where neither performance nor fairness can be improved without harming the other. In other words, they lie on the "Pareto frontier" or the "Skyline" of possible solutions, where you can't get a better outcome in any of the objectives (performance or fairness) without making the other worse. The term Skyline is used to describe a set of Pareto-optimal solutions, where the best options are those that are not dominated by others.

3.1 Pareto Optimality and In the Context of Fairness:

Pareto optimality plays a crucial role in understanding the trade-offs between fairness and performance in machine learning. Consider multiple groups g_i to g_k, where the performance of an algorithm for group g_i is denoted as $Per(i)$. Let $Per^u(i)$ represent the performance of an unfair algorithm for g_i and $Per^f(i)$ represent the performance of a fair algorithm for g_i. If for all groups g_i, the condition

$$Per^u(i) \geq Per^f(i), \quad \forall i$$

holds, then the fair solution is said to be **dominated** by the unfair solution. This means that the unfair algorithm performs strictly better or at least as well for every group, making the fair solution suboptimal in terms of performance.

3.2 Pareto Optimal Solutions:

Pareto optimality refers to solutions that are not dominated by any other solution. A Pareto-optimal solution is one in which no further improvement can be made in any objective (like fairness or performance) without making the other worse. Pareto optimality suggests that achieving fairness and performance simultaneously requires compromises.

Causality - Chapter 5:

4 Causality: A Double-Edged Sword

Causality is a powerful tool for understanding the underlying reasons behind disparities in outcomes, particularly when evaluating fairness and equity. It enables researchers and practitioners to go beyond simple observations and explore whether differences in outcomes are due to legitimate causes or hidden biases.[2] In fields like social sciences, medicine, and machine learning, causal reasoning plays a key role in identifying the root causes of inequality, enabling more informed decision-making.

However, the strength of causal analysis also makes it a double-edged sword. While it can reveal deep insights, it can also be misleading if used improperly. Causal models rely on specific assumptions, and flawed assumptions can lead to deceptive conclusions[2]. A central concept in causal reasoning is counterfactual reasoning, where we ask questions like "What would have happened if things were different" Counterfactuals allow us to compare the actual world with hypothetical alternatives, helping us distinguish true causal relationships from mere correlations [3].

5 Correlation vs. Causation

A fundamental principle in data analysis and causal inference is that correlation does not imply causation. Just because two variables are statistically related does not mean that one causes the other. Misinterpreting correlation as causation can lead to flawed conclusions and misguided decisions[2].The distinction between observation and action is essential here. Observational data captures the natural co-occurrence of events without any intervention. Causal inference, on the other hand, focuses on the impact of deliberate actions or interventions. Using methods like the do-operator in structural causal models, researchers can simulate scenarios where specific variables are manipulated to observe their direct effects. This helps determine whether changing one factor actually causes a change in another, rather than simply being associated with it.

6 Modeling Causality: The Need for Structured Approaches

To effectively analyze causality, researchers use formal frameworks that go beyond simple statistical methods. One of the most robust approaches is the *Structural Causal Model* (SCM), which provides a systematic way to represent and study causal relationships. By providing a structured way to model the data-generating process, SCMs help researchers avoid common pitfalls in causal analysis, such as misinterpreting spurious correlations or falling victim to selection bias.

7 Structural Causal Model (SCM): The Strongest Model for Causality

7.1 Informal Definition

A Structural Causal Model (SCM) can be understood through the example of choosing between two routes to a destinationwith probability od 1/2. Each route has a 1/3 chance of an accident, influenced by external factors like traffic or weather. A causal graph represents this scenario, where nodes represent variables (e.g., "Route Chosen," "Traffic," "Accident") and edges show causal connections. The graph illustrates how route choice impacts accident risk, while external factors affect both traffic and accident likelihood. This demonstrates how causal models capture dependencies and enable predictions, highlighting the difference between correlation and causality.

7.2 Formal Definition

A Structural Causal Model (SCM) provides a rigorous mathematical framework to represent causality.[2] It consists of:

- A set of random variables $X_1, X_2, \ldots, X_d$ that represent the different elements of the system being studied.
- A set of structural equations that define how each variable is generated. Specifically, each variable X_i is determined by a function:

$$X_i = F(P_i, U_i)$$

  where:

  - P_i represents the parents of X_i—the direct causes of X_i within the causal graph.
  - U_i is a noise term accounting for external influences or randomness not explicitly modeled.

This setup allows SCMs to generate samples from the joint probability distribution of the variables, but unlike traditional statistical models, SCMs retain information about the causal structure. This means researchers can perform interventions—such as setting a variable to a fixed

value using the do-operator (e.g., $do(X = x)$)—and observe how the distribution of other variables changes as a result.

8 Causal Graphs

A Structural Causal Model (SCM) can be effectively represented using a Directed Acyclic Graph (DAG), where nodes represent variables, and directed edges denote causal relationships between these variables. DAGs provide a structured way to visualize and analyze how different factors influence each other within a system. In such graphs, the absence of cycles ensures a clear directional flow of causality, which aligns with the principle that causes precede effects.

8.1 Example: The Ice Cream Shop and Confounding Variables

Consider an ice cream shop where higher electricity bills correlate with increased ice cream sales. At first glance, one might assume that electricity usage drives sales. However, the true confounder is temperature.[3] Hot days lead to more air conditioning use (higher bills) and more ice cream sales.

This example highlights why simple correlations can be misleading and how causal reasoning helps identify hidden variables driving observed outcomes. The causal relationships can be depicted as:

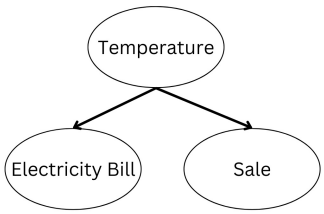

Figure 1: Causal Graph for the ice cream example

This representation 1 highlights the importance of recognizing confounding variables—like temperature, that simultaneously influence multiple outcomes, helping to avoid incorrect causal conclusions.

8.2 Example: Simpson's Paradox

Simpson's Paradox shows how aggregated data can hide true causal relationships. In the 1973 UC Berkeley admissions, data suggested gender bias—44% of men were admitted versus 35% of women. Yet, when broken down by department, most favored women. The overall bias resulted from women applying more to competitive departments with lower acceptance rates [2].

This paradox underscores the importance of causal analysis to avoid misinterpreting aggregated data and to identify hidden factors influencing outcomes. This nuanced relationship can be captured using a causal graph:

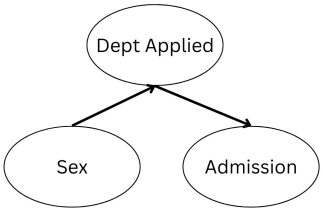

Figure 2: Causal Graph for the Simpson's Paradox

The causal graph 2 clarified that the real driver of the disparity was the choice of the department, not the direct gender bias in admissions.

8.3 Observations About Causal Graphs

- **No Cycles Allowed:** In causal graphs, cycles are not allowed. A cycle would imply a feedback loop where a variable could indirectly cause itself, violating the fundamental principle of causality where causes precede effects.
- **Topological Sorting:** Since cycles are absent, variables in a causal graph can be topologically sorted - ordered in a sequence. This ordering reflects the natural causal structure and allows for step-by-step reasoning about interventions and outcomes.
- **DAGs as Bayesian Networks:** Causal graphs are a subclass of Bayesian networks, which are probabilistic graphical models that represent conditional dependencies between variables. In causal graphs, the edges have a specific interpretation as causal relationships, not just statistical associations.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. *A reductions approach to fair classification*, 2018.

[2] Solon Barocas. *Fairness and Machine Learning*. n.d. Fairness and Machine Learning.

[3] Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). *A survey of learning causality with data: Problems and methods*. ACM Computing Surveys (CSUR), 53(4), 1-37.