| DS603: Robust Machine Learning | Autumn 2025 |
|---|---|
| Lecture 1 — 27th November | |
| *Lecturer: Arjun Bhagoji* | *Scribe: Your Name* |

## 1.1 Introduction: The Bias Problem

Artificial Intelligence (AI) systems learn from vast amounts of digitized historical data (books, news, records). Consequently, they risk "baking old attitudes into new technology". The lecture introduced several domains where this algorithmic bias manifests with severe consequences.

### 1.1.1 Banking and Lending

- **Disparate Denial Rates:** In mortgage applications, algorithms have been shown to deny Black applicants at disproportionately high rates. One report noted that 80% of Black mortgage applicants were denied.
- **Controlled Comparison:** Even when controlling for credit scores, disparities persist. Applicants of color were found to be 20% to 120% more likely to be denied than White applicants with comparable credit scores.
- **Selection Rate Disparity:** Mathematically, if we denote the decision as $F(x)$, this observation implies:
$$P[\text{Loan Denial} \mid \text{Score} = C, \text{Black}] > P[\text{Loan Denial} \mid \text{Score} = C, \text{White}]$$

### 1.1.2 Healthcare

- **Skin Cancer Detection:** Machine learning models for detecting skin cancer often suffer from "bias in, bias out" due to the underrepresentation of diverse skin types in training data.
- **Chest Radiographs:** Algorithms applied to chest X-rays have shown "underdiagnosis bias" in underserved populations.
  - *False Positive Rate (FPR) Disparity:* Higher FPR for women compared to men ($\Delta > 0$). This leads to women being falsely flagged as healthy or receiving unnecessary treatment.
  - *False Negative Rate (FNR) Disparity:* Conversely, different groups may experience higher false negative rates, leading to missed diagnoses and lack of timely care.

## 1.2 Formalizing Algorithmic Bias

### 1.2.1 Notation and Setup

We consider a supervised classification setting:

- $X \in \mathbb{R}^d$: Task-specific feature vector (e.g., credit history, medical vitals).

- $Y \in \{0,1\}$: True binary class label (e.g., default/repay, disease/healthy).
- $Z \in \{0,1\}$: Protected group attribute (e.g., race, gender).

The goal is to learn a classifier $F : \mathbb{R}^d \to \{0,1\}$ that minimizes an empirical loss $\omega(f;\mathscr{D})$ over the dataset $\mathscr{D} = \{(x_i, y_i, z_i)\}_{i=1}^N$:

$$F = \arg\min_f \omega(f;\mathscr{D})$$

Common loss functions include:

- **0–1 Loss:** $\omega(f;\mathscr{D}) = P_{(X,Y)\sim\mathscr{D}}[f(X) \neq Y]$.
- **Log-Loss:** $\omega(f;\mathscr{D}) = \mathbb{E}[-Y\log(f(X)) - (1-Y)\log(1-f(X))]$.

### 1.2.2 Fairness Metrics

Fairness is defined by statistical independence or separation criteria.

#### Demographic Parity (Independence)

Requires the prediction to be independent of the protected attribute.

$$P[F(X) = 1 \mid Z = 1] = P[F(X) = 1 \mid Z = 0]$$

The **Demographic Disparity** is $\Delta_{DP} := |P[F = 1 \mid Z = 1] - P[F = 1 \mid Z = 0]|$.

- *Note:* Satisfying this may require different thresholds for different groups if the underlying score distributions differ (e.g., credit score CDFs differ by race).

#### Equalized Odds (Separation)

Requires conditional independence of the prediction and protected attribute given the true label ($F \perp Z \mid Y$). This encompasses:

- **False Positive Parity:** $P[F = 1 \mid Y = 0, Z = 1] = P[F = 1 \mid Y = 0, Z = 0]$.
- **False Negative Parity:** $P[F = 0 \mid Y = 1, Z = 1] = P[F = 0 \mid Y = 1, Z = 0]$.

#### Calibration Parity

Requires that for any predicted probability score $s$, the probability of the positive class is the same across groups:

$$P[Y = 1 \mid F(X) = s, Z = 0] = P[Y = 1 \mid F(X) = s, Z = 1]$$

## 1.3 Fairness-Constrained Optimization

### 1.3.1 The Problem

We formulate fairness as a constrained optimization problem:

$$\begin{aligned} \min_{f} \quad & \omega(f;\mathscr{D}) \\ \text{subject to} \quad & \Delta_h(f) \leqslant \varepsilon \end{aligned} \tag{1}$$

where $\Delta_h(f)$ is a fairness metric (e.g., demographic disparity) and $\varepsilon$ is a tolerance.

### 1.3.2 Optimization Challenges

1. **Non-Convexity:** Fairness constraints on discrete outputs are often non-convex and difficult to optimize directly.
2. **Saddle Points:** Solving the Lagrangian formulation $L(f,\lambda) = \omega(f) + \lambda(\Delta_h(f) - \varepsilon)$ requires solving a min-max problem: $\min_f \max_{\lambda \geqslant 0} L(f,\lambda)$.

### 1.3.3 Relaxations (Zafar et al. 2017)

To make the problem tractable, we can relax the independence constraint to a covariance constraint on the decision boundary weights $w$ (where $f(x) = \text{sign}(w^T x)$):

$$\text{Cov}(w^T X, Z) \approx 0 \implies |w^T \mathbb{E}[X(Z - \bar{Z})]| \leqslant \varepsilon$$

This linear constraint allows the use of standard convex solvers.

## 1.4 The Price of Fairness

Imposing fairness constraints often results in a reduction in utility (accuracy) on the training data. This trade-off is formalized as:

$$\text{Price of Fairness} = \omega(F_{\text{fair}};\mathscr{D}) - \omega(F_{\text{unconstrained}};\mathscr{D})$$

While usually positive ($> 0$), this "price" reflects performance on the specific dataset and does not account for broader societal costs or long-term equality.

## 1.5 Normative and Practical Considerations

Selecting a fairness metric is not merely a technical choice but a normative one guided by:

### 1.5.1 Legal Guidelines

- **Indian Constitution (Art. 14):** Guarantees equality but permits "reasonable differential treatment" if it has a rational relation to the objective.
- **Affirmative Action:** Policies that explicitly encode fairness constraints to address inequality.

### 1.5.2 Philosophical Frameworks

The lecture mapped fairness metrics to political theories:

- **Utilitarianism:** Fairness should reduce errors and improve performance for *all* groups.
- **Egalitarianism:** Fairness should ensure equal rights and opportunities for all.
- **Rawls' Principle (Maximin):** Fairness should improve the standing of the worst-off group.
- **Anti-subordination:** Fairness should address historical inequalities, even if it requires asymmetric treatment (corrective justice).
- **Non-arbitrariness:** Avoidance of arbitrary decision-making.

### 1.5.3 Participatory Methods

Beyond theory, practitioners should use "Preference Elicitation" to understand stakeholders' specific fairness expectations and acceptable fairness-utility trade-offs.

- **Legal guidelines:** Provide concrete constraints and requirements (for example, anti-discrimination law may prohibit certain disparate treatment or require specific forms of reasonable accommodation). Legal rules translate normative commitments into implementable tests and remedies and thus often place hard bounds on acceptable fairness choices.
- **Moral and political theories:** Offer philosophical justifications that shape which fairness desiderata are appropriate in a context. Theories commonly invoked include:
  - **Utilitarianism:** Prioritizes aggregate welfare—under this view, a fairness intervention is desirable if it reduces total errors or improves overall outcomes across all groups.
  - **Egalitarianism:** Emphasizes equal rights and opportunities—policies are preferred when they produce equal treatment or equal access to desirable outcomes.
  - **Non-arbitrariness:** Emphasizes procedural regularity and reasons—fairness requires that decisions not be arbitrary and that similarly situated individuals be treated consistently.
  - **Rawls' Maximin (Difference Principle):** Favors interventions that improve the position of the worst-off; fairness constraints are justified when they raise the standing of the least advantaged group.
  - **Anti-subordination:** Focuses on remedying historical and structural disadvantages; it can justify asymmetric treatment (e.g., affirmative action) if that corrects entrenched inequality.
- **Participatory and empirical methods:** Eliciting preferences from affected individuals and communities helps operationalize normative choices. "Preference elicitation" can be done at the individual level (surveys, conjoint analysis) and the community level (deliberative workshops, stakeholder panels). These methods reveal which trade-offs stakeholders find acceptable and surface context-specific harms that abstract metrics might miss.

### 1.5.4 Putting It Together: Choosing Constraints in Practice

Choosing a fairness constraint for a deployed system typically follows a pragmatic, multi-step process:

1. **Identify legal requirements:** First determine any hard constraints imposed by law or regulation in the deployment jurisdiction; these may rule out some options entirely.
2. **State normative commitments:** Make explicit the normative stance the project adopts (e.g., prioritizing worst-off groups, equalizing error rates, or maximizing aggregate utility). This drives which mathematical constraint to prefer.
3. **Elicit stakeholder preferences:** Run participatory exercises with affected communities, domain experts, and operators to learn which fairness–utility trade-offs are acceptable.
4. **Assess data and measurement limits:** Check whether the available data permit reliable measurement of both outcomes and protected attributes; measurement error can undermine many fairness constraints.
5. **Model and compare interventions:** Evaluate candidate constraint approaches (demographic parity, equalized odds, calibration, targeted remedies) on representative data and compare their impacts on accuracy, subgroup harms, and downstream outcomes.
6. **Choose a deployment strategy:** Decide whether to use pre-processing, in-processing, post-processing, or socio-technical remedies (policy changes, human-in-the-loop workflows). Consider monitoring plans and rollback criteria.
7. **Document and monitor:** Record the rationale for the chosen constraint, the stakeholder input, and the expected trade-offs. Monitor model performance and fairness metrics in production and re-run stakeholder consultation periodically.

| Fairness Metric | Pro | Con |
|---|---|---|
| Demographic Parity | Simple to enforce | Ignores base rates |
| Equalized Odds | Error-rate focused | Requires labels |
| Calibration | Probabilistic | Hard to satisfy jointly |

**Table 1.1.** Comparison of different fairness metrics.