**Pareto Efficient Fairness for Skewed Subgroup Data**

Ananth Balashankar, Alyssa Lees, Chris Welty, Lakshminaryanan Subramanian
ananth@nyu.edu, {alyssalees, welty}@google.com, lakshmi@nyu.edu

ML can amplify societal biases

Perfect Fairness and Accuracy are at odds

- **Theorem**: If the class probabilities of the target (Y) and sensitive features (S) are aligned, there is an unavoidable degradation in accuracy owing to the fairness requirement [Menon and Williamson, 2018]

- Examples of Perfect Fairness Requirements (assuming Y is binary):
  - Demographic Parity: P(Ŷ=1|S=m) = P(Ŷ=1|S=n) [Calders and Verwer, 2010]
  - Equality of Opportunity: P(Ŷ=1|Y=1, S=m) = P(Ŷ=1|Y=1, S=n) [Hardt et al., 2016]
  - Equality of Odds: P(Ŷ=ŷ|Y=y, S=m) = P(Ŷ=ŷ|Y=y, S=n); ŷ, y ∈ {0,1}, ∀m, n ∈ S, m ≠ n [Hardt et al., 2016]

Existing Approaches

- Approximate Fairness with Lagrangian Constraints [Zhao et al., 2017]
  - Augments a fairness penalty term to the cross entropy loss
  - The penalty factor λ is hard to fine-tune by the domain expert
- Adversarial Multi-Task Learning [Beutel et al., 2017]
  - Learn the target and ensure that the sensitive feature is not learnt through negative gradients of multi-task learning
  - No control given to domain expert to control the trade-off meaningfully

What we want in many cases...

- Policies like affirmative action and UN Sustainable Development goals aim to improve performance of protected groups to meet the levels of the highest performing groups [Foster and Vohra, 1992]
- In skewed subgroup datasets, there might be an opportunity to choose performance for all groups that are better than the best perfectly fair one
- We want to give the domain expert the ability to search for Pareto-Efficient performance within a Fairness bound

Pareto-efficient fairness is better

- Pareto-Efficient: Set of points for which there does not exist another point, which is better performing across all sensitive groups
- Fairness: The deviation from each group’s Pareto-optimal point is distributed equitably among groups

Pareto efficient bias mitigation

\[ L_p = L_{ce} + \lambda(\alpha\|E_G\|_1 + (1 - \alpha)\alpha^2(G)) \]

- G: set of sensitive groups, D: dataset, D_g: data of group g ∈ G
- for \( g \in G \) do
  - \( M_g = \arg \min L_{ce}(D_g) \)
  - \( f_{opt_g} = \text{eval}(M_g, D_g) \)
  - \( f_g = 0 \)
- end for
- while \( \exists g \in G, f_g = 0 \) ∨ \( f_g > f_{opt_g} \) do
  - \( f_{opt-p} = \max(f_g, f_{opt-g}), \forall g \in G \)
  - \( M = \arg \min L_{p}(D) \)
  - \( f_g = \text{eval}(M, D_g), \forall g \in G \)
- end while

Evaluation on UCI Census

- **data**
  - 14 demographic features like age, education, occupation, etc from 1994 census
  - Predict if income is > 50K or not
  - Sensitive variables assumed are race and gender
- Pareto Efficient loss is minimized, while achieving best overall accuracy

**Table 1. UCI Adult dataset with bias mitigation algorithms**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>TPR</th>
<th>TNR</th>
<th>Discrepancy</th>
<th>Pareto Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no bias loss)</td>
<td>0.830</td>
<td>0.733</td>
<td>0.247</td>
<td>0.190</td>
<td>0.016</td>
</tr>
<tr>
<td>Minimize Discrepancy</td>
<td>0.619</td>
<td>0.283</td>
<td>0.712</td>
<td>0.167</td>
<td>0.133</td>
</tr>
<tr>
<td>Adversarial Loss</td>
<td>0.648</td>
<td>0.224</td>
<td>0.769</td>
<td>0.226</td>
<td>0.077</td>
</tr>
<tr>
<td>Pareto-Efficient Loss</td>
<td>0.678</td>
<td>0.145</td>
<td>0.850</td>
<td>0.250</td>
<td>0.000</td>
</tr>
</tbody>
</table>

- Pareto Efficient loss ensures best accuracy across all subgroups too!

**Table 2. Subgroup performance on UCI Adult dataset**

<table>
<thead>
<tr>
<th>Model</th>
<th>Subgroup 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Pareto Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no bias loss)</td>
<td>0.890</td>
<td>0.883</td>
<td>0.818</td>
<td>0.784</td>
<td>0.016</td>
</tr>
<tr>
<td>Minimize Discrepancy</td>
<td>0.853</td>
<td>0.856</td>
<td>0.806</td>
<td>0.778</td>
<td>0.153</td>
</tr>
<tr>
<td>Adversarial Loss</td>
<td>0.882</td>
<td>0.872</td>
<td>0.824</td>
<td>0.780</td>
<td>0.077</td>
</tr>
<tr>
<td>Pareto-Efficient Loss</td>
<td>0.935</td>
<td>0.915</td>
<td>0.844</td>
<td>0.797</td>
<td>0.000</td>
</tr>
<tr>
<td>Subgroup Pareto Frontier</td>
<td>0.934</td>
<td>0.894</td>
<td>0.815</td>
<td>0.783</td>
<td>N/A</td>
</tr>
</tbody>
</table>

“Men Also Like Shopping Reducing Gender Bias Amplification using Corpus-level Constraints” [Zhao et al., 2017]

“Gender, race and the presentation of acute coronary syndrome and serious cardiopulmonary diagnosis in ED patients with chest pain” [Alabban et al., 2017]