Decentering Imputation: Fair Learning at the Margins of Demographics

Evan Dong and Cyrus Cousins

Brown University Department of Computer Science

ICML International Conference On Machine Learning

July 2022

QUEER in AI

When do we lack demographic labels?

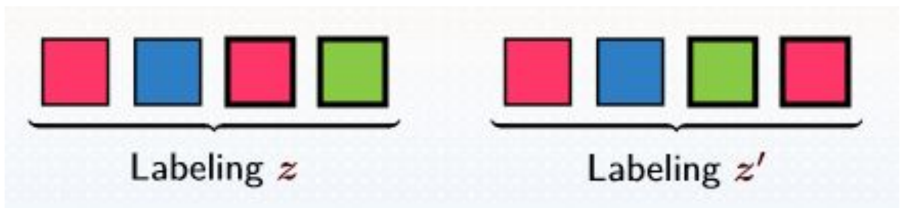
- Data collection or access may be regulated
  - E.g., medically relevant disability information
  - "Refusing Research," data sovereignty
- Categorization may be flawed, incomplete, or politicized
  - E.g., binary gender categories
- Privacy concerns or non-response may limit availability
  - "Prefer not to respond"
  - Limited disclosure from dataset owners to model creators

Flaws of Imputation

- Proxies are imperfect
  - Statistical biases, few guarantees
- Replicates ecological fallacy
  - Expecting group-level trends in individuals
- Prediction centers the *mean* instead of the *margins*
  - Fails the less represented (bell hooks)
  - E.g., low-income Asian communities
- Imposes structural assumptions, patterns
  - E.g., gender essentialism in gender recognition
  - Correlations built on historical inequality

Setting

- Task dataset (X → Y)
  - Unknown *z*
- Auxiliary dataset (X → Z)
  - *y* data optional
- Idea: project *possible z* distributions onto task data
- Statistics: similar populations mean similar *z* distributions



- Distribution transferred across the whole dataset
- Refocus from mean to confidence interval
  - Uncertainty from *individuals* to *groups*
- How can we provide fairness guarantees?
  - Find "worst-case" groups for a hypothesis
  - Bounded error by construction

Rawlsian Fairness

- John Rawls's Maximin Principle
  - "Maximize the welfare of the least well-off"
- In machine learning:
  - Minimize greatest conditional loss across groups

$$\operatorname*{argmin}_{\theta\in\Theta} \max_{j\in\mathcal{Z}} \mathop{\mathbb{E}}_{x,y}\left[\ell(h_\theta(x),y) \mid Z=j\right]$$

Constructing an Adversary

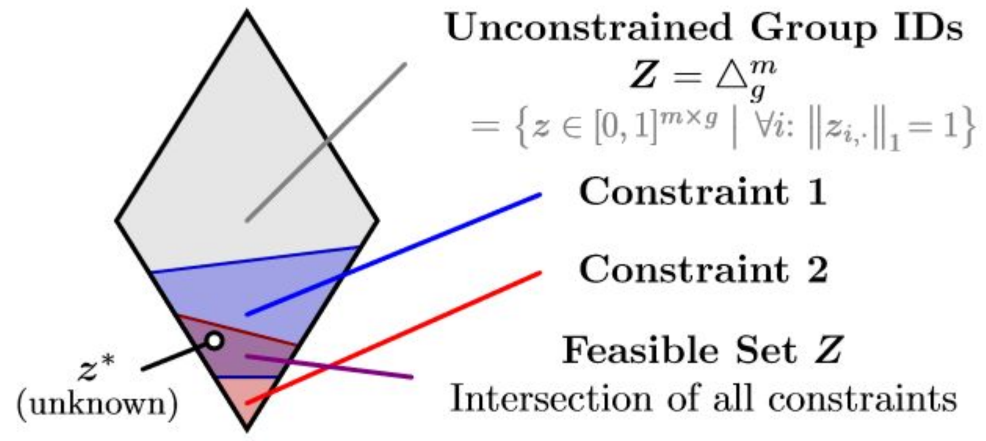
- Find inequality-maximizing distribution of group memberships across the task data
- Limited to feasible set ***Z*** over task demographic labelings

$$\max_{z\in \boldsymbol{Z}} \max_{j\in\mathcal{Z}} \frac{\sum_{i=1}^{m} z_{i,j}\,\ell(h_\theta(x_i),y_i)}{\sum_{i=1}^{m} z_{i,j}}$$

- Discrete 0-1 labels = integer programming (NP-hard)
  - Relax to continuous *z* simplex
- Linear-fractional program
  - Solvable as an efficient linear program (Charnes-Cooper transformation)
  - *O(mg)* variables, *O(m + c)* constraints

Defining a Feasible Set

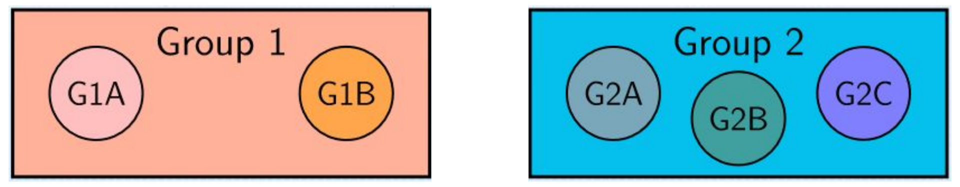
- If all labelings are possible, fairness is unachievable
- What are probable sets of demographic labels?
  - What possibilities are *ruled out* with prior knowledge?
- Create constraints from auxiliary data



- **Size of feasible set = strength of adversary**
- More constraints lead to narrower possibilities
  - Greater knowledge of group-level characteristics
  - Never make assumptions about *individuals*
  - Customizable

Creating Constraints

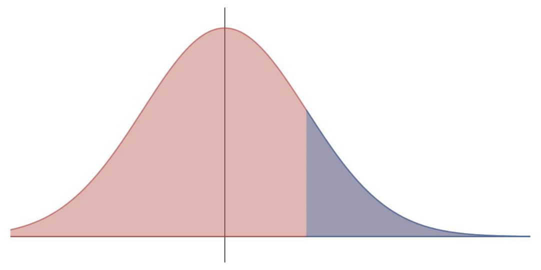
- Variety of possible sources
  - Census statistics, demographic surveys, etc.
  - Clusters, estimating group frequencies
  - Auxiliary pre-trained group predictors (X → Z)
    - Constrain *accuracy* of predictors on task data
  - Hierarchical relationships: groups and subgroups



Tail Bounds

$$\mathbb{P}\left(|\bar{u}-\mu| > \varepsilon\right) \le \delta\ .$$

- Sample mean within ε of expectation at probability $1-\delta$
  - E.g., with prior estimate that 60% of the population is *A*, 95% of the time, 58-62% of a sample will be *A*
- Bounds come from both datasets
- Auxiliary dataset
  - Rademacher Averages
  - Bousquet's Inequality
- Task-dependent:
  - Hoeffding's Inequality

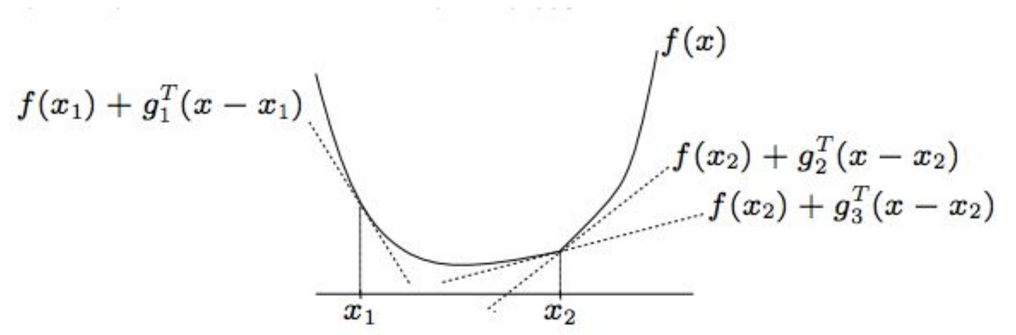


$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} u_i - \mathbb{E}[u]\right| \ge \sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) \le \delta$$

*(for unit range)

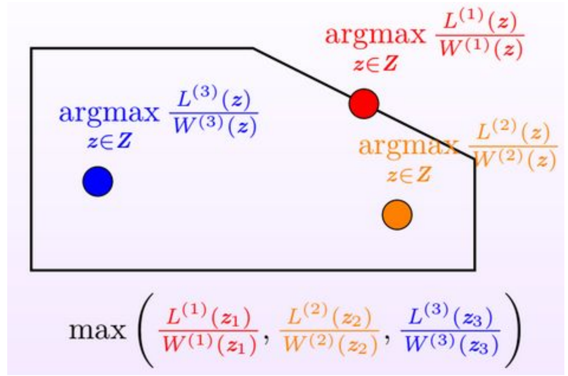
Subgradient Method

- Works for any convex loss
- Subgradient is easily to calculate
- Converges within $O(1/\varepsilon^2)$ steps



Efficient Training

- Repeated large linear programs: costly
- Deploy after naive, fairness-agnostic training
- Cache prior solutions as approximations
  - Lipschitz bound on loss changes
  - Only solve for each group when needed
  - Prior solutions are close approximations



Advantages over Imputation

- Avoid assumptions about individuals
  - Ecological Fallacy, Aggregation Bias
  - Reproducing processes of racialization, gendering, etc.
- Limit worst-case error, not unfounded expectation
- Better define "the least well-off" as per Rawls
- More explainable training
  - demographic prediction errors do not compound
- Can combine sources of information

Comparison to Prior Work

- Distributionally Robust Learning
  - Optimization over time
- Adversarially Reweighted Learning
  - Requires highly distinguishable trends in groups
- Explicitly defining protected groups can be *valuable*

Future Work

- Experimental validation
  - Work-in-progress
- Explore privacy guarantees
  - Differentially private adversary constraints?
  - Protect against database reconstruction
- Faster verification of convergence
  - Defining "good initialization"
  - Optimize learning rate
- Framework for combining constraint sources