Minimax Group Fairness: Algorithms and Experiments

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth

University of Pennsylvania, Amazon AWS AI

Motivation

- Machine learning researchers and practitioners have often focused on achieving group fairness with respect to protected attributes (race, gender, ethnicity, etc.)
- **Equality of error rates** is one of most intuitive and well-studied group fairness notions
- But in practice, equalizing error rates and similar notions may require **artificially inflating error** on easier-to-predict groups and **may be undesirable** for a variety of reasons
- There are many social applications of machine learning in which most/all of the targeted population is disadvantaged
- Might be interested in ensuring predictions are roughly equally accurate across racial groups, income levels, geographic location, etc
  - But, if this can only be achieved by raising lower group error rates, then we have worsened overall social welfare
- Therefore, might be preferable to consider the alternative fairness criterion of **minimax group error**, recently proposed by [Martinez, 2020]
  - Seek not to equalize error rates, but to minimize largest group error rate, making sure that **the worst-off group is as well-off as possible**

Contributions

1. Propose two algorithms, both two player zero-sum games:
   1.1 MINIMAXFAIR: Finds a minimax group fair model from a given statistical class
   1.2 MINIMAXFAIRRELAXED: Finds a model that minimizes overall error subject to the constraint that all group errors must be below a predetermined threshold
   - **Navigates tradeoffs** between a relaxed notion of minimax fairness and overall accuracy
2. Prove that both algorithms converge and are oracle efficient.
3. Show how our framework can be extended to handle different types of error rates, such as false positive (FP) and false negative (FN) rates, as well as overlapping groups
4. Provide a thorough experimental analysis of our two algorithms under different prediction regimes

Mathematical Framework

- Consider pairs of dependent and independent variables $(X_i, y_i)_{i=1}^n$ divided into K groups $\{G_1, ..., G_K\}$, class H of (potentially unfair) mixtures of statistical models, with loss function L and average group loss ϵ_k for some $h \in H$:

$$\epsilon_k(h) = \frac{1}{|G_k|} \sum_{(x,y)\in G_k} L(h(x), y)$$

1. In pure minmax problem, goal is to find a mixed strategy h^* that minimizes the maximum error rate over all groups:

$$h^* = \text{argmin}_{h \in H}\{\max_k \epsilon_k(h)\} \tag{1}$$

2. In relaxed version, specify max group error γ and model that minimizes overall population error while staying below the maximum group error threshold:

$$\begin{aligned} &\underset{h \in H}{\text{minimize}} && \epsilon(h) \\ &\text{subject to} && \epsilon_k(h) - \gamma \le 0, k = 1, \dots, K \end{aligned} \tag{2}$$

Algorithmic Formulation: Two Player Zero-Sum Game

Can recast both problems as a zero-sum game between a (L)earner and a (R)egulator:

- At each round t, there is a weighting over groups determined by R
- L (best) responds by computing model h_t to minimize the weighted prediction error
- R updates group weights using exponential weights/gradient ascent with respect to group errors achieved by h_t
- L's final model M is uniform distribution over all of $h_t$âs produced

MINIMAXFAIR

Algorithm 1: MINIMAXFAIR

Input: $\{x_i, y_i\}_{i=1}^n$, adaptive learning rate η_t, populations G_k with relative sizes $p_k = \frac{|G_k|}{n}$, iteration count T, loss function L, model class H

Let $\epsilon_k(h) = \frac{1}{|G_k|}\sum_{(x,y)\in G_k} L(x,y)$

Initialize $\lambda_k = p_k \; \forall k$

for $t = 1$ ***to*** T **do**

  Find $h_t =_{h\in H} \sum_k \lambda_k * \epsilon_k(h)$

  Update each $\lambda_k = \lambda_k * exp(\eta_t * \epsilon_k(h_t))$

end

Output: Uniform distribution over set of models $h_1, ..., h_T$

MINIMAXFAIRRELAXED

Algorithm 2: MINIMAXFAIRRELAXED

Input: $\{X_i, y_i\}_{i=1}^n$, adaptive learning rate η_t, populations G_k with relative sizes $p_k = \frac{|G_k|}{n}$, iteration count T, loss function L, model class H, maximal group error γ

Let $\epsilon_k(h) = \frac{1}{|G_k|}\sum_{(x,y)\in G_k} L(x,y)$

Initialize $\lambda_k = 0 \; \forall k$

for $t = 1$ ***to*** T **do**

  Find $h_t = argmin_{h\in H} \sum_j (p_k + \lambda_k) * \epsilon_k(h)$

  Update each $\lambda_k = \max(\lambda_k + \eta * (\epsilon_k(h_t) - \gamma), 0)$

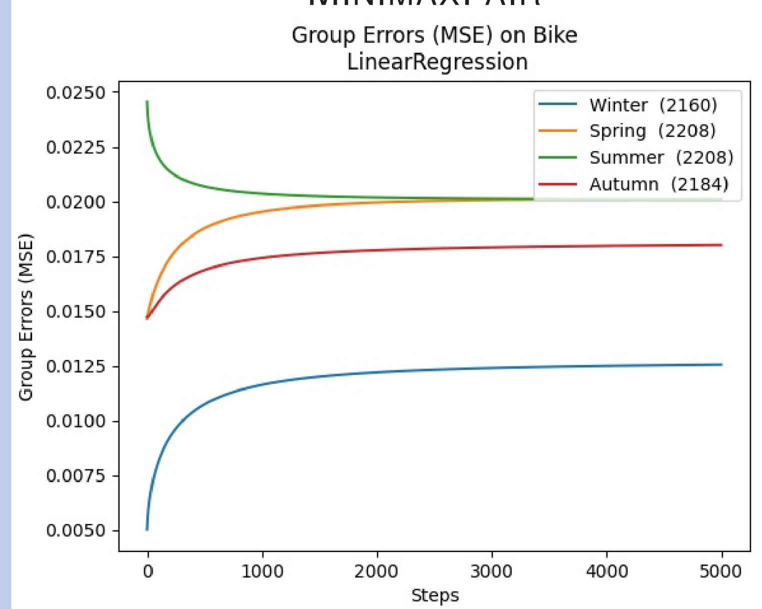
end

Output: Uniform distribution over set of models $h_1, ..., h_T$

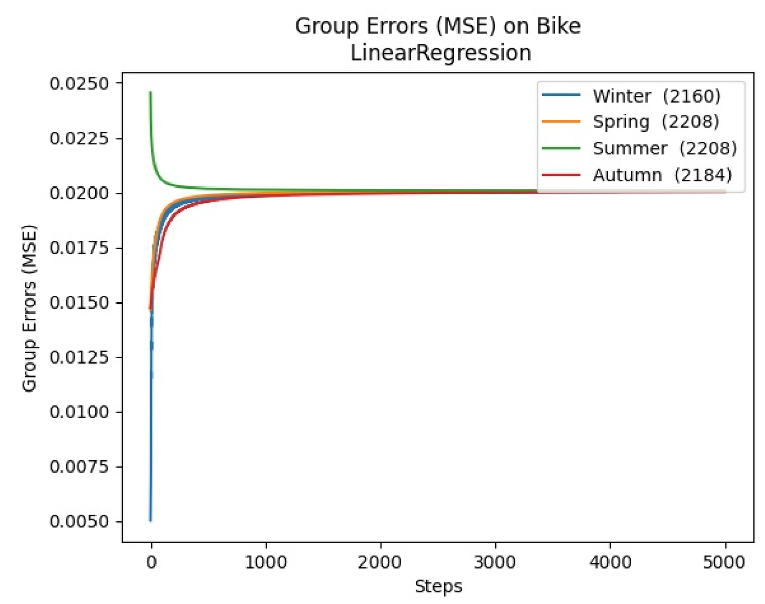
MINIMAXFAIR vs. Equal Errors Regression

Comparison of Minimax and Equal Error Solutions on Seoul Bike Dataset

MINIMAXFAIR



Equal Errors

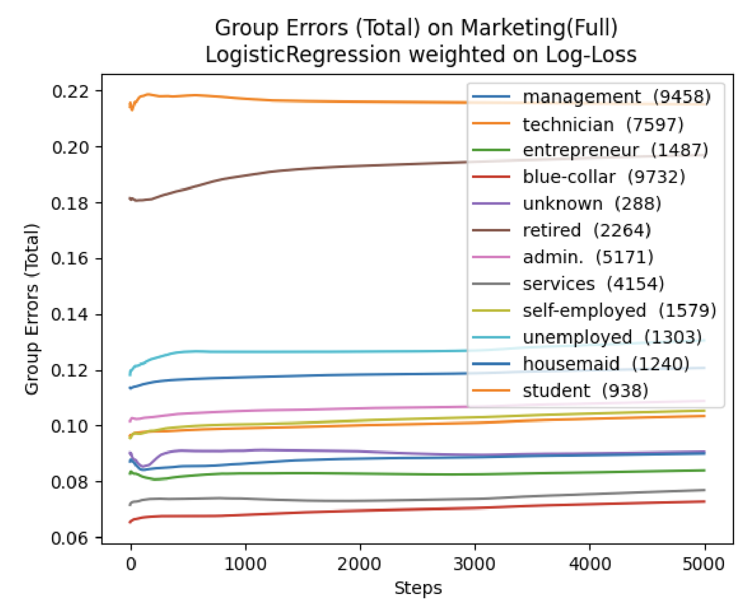


Dataset: Public bikes rented at each hour in Seoul Bike sharing system
Label: Rented bikes (normalized), *Group*: Season

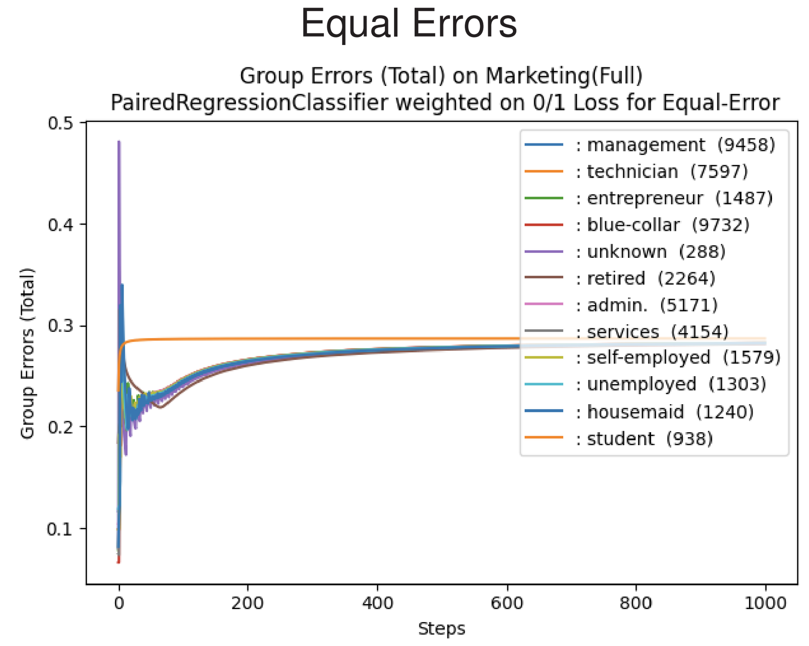
MINIMAXFAIR vs. Equal Errors Classification

Comparison of Minimax and Equal Errors on Marketing Dataset

MINIMAXFAIR



Equal Errors

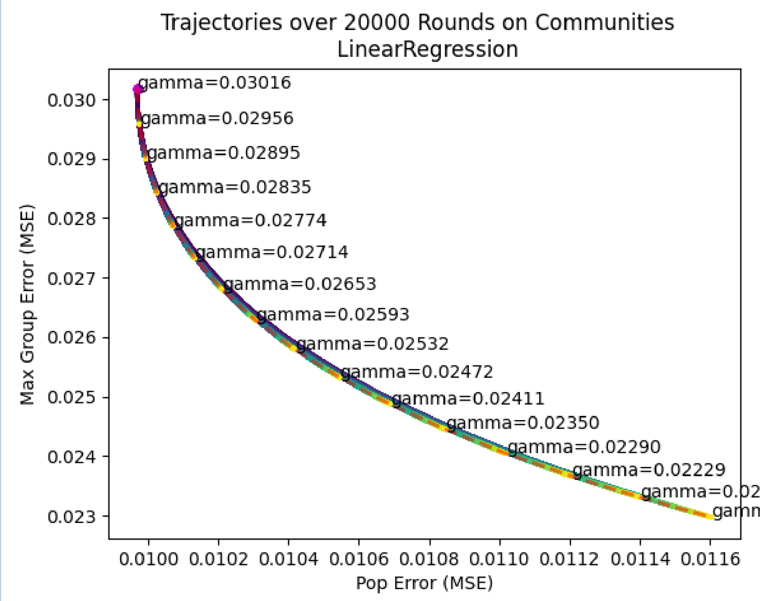


Dataset: Direct marketing campaigns (phone calls) of a Portuguese bank
Label: client subscribes term deposit, *Group*: Job

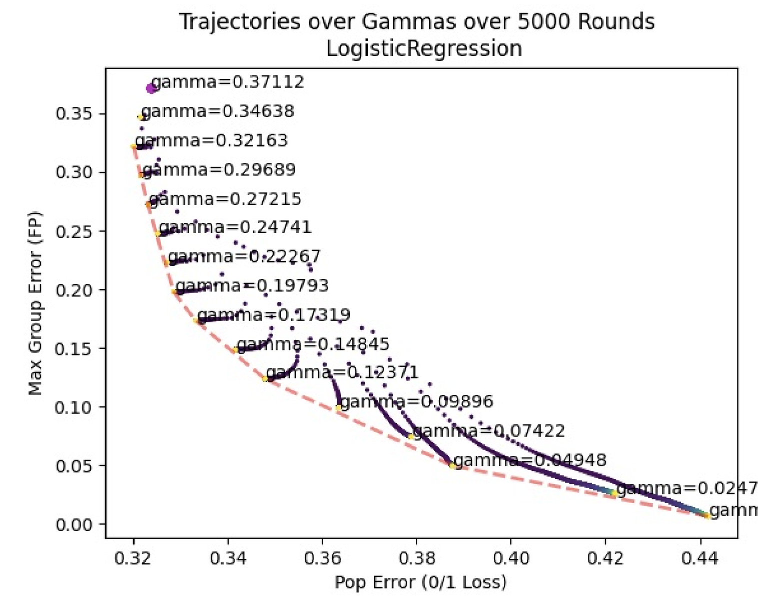
Fairness Accuracy Tradeoff with MINIMAXFAIRRELAXED

Fairness Accuracy Tradeoff Curves

Linear Regression on Communities Dataset



Classification (FP) on COMPAS Dataset



Communities and Crime: US Communities, 1990 - 1995, *Label*: Violent crimes per population, *Group*: Race
COMPAS: Arrest data from Broward County, Florida, *Label*: Two year recidivism, *Groups*: Race, sex

Generalization Results

- With probability $1 - \delta$, generalization gap per group bounded by

$$O\left(\sqrt{\frac{\log\frac{1}{\delta} + d\log n_i}{n_i}}\right)$$

  where d is VC dimension of class H, and n_i is sample size of group i
- Generalization gap for *minimax group* is bounded by

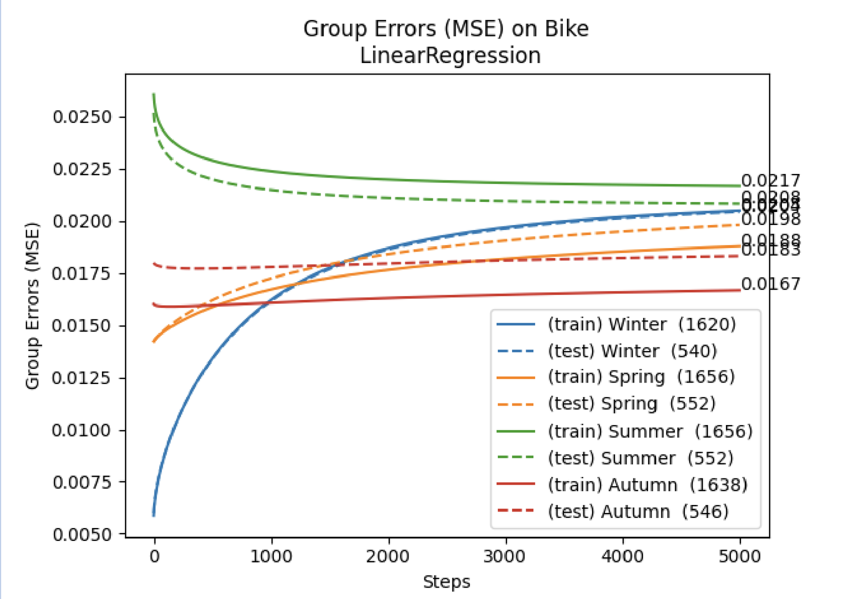
$$O\left(\max_i \sqrt{\frac{\log\frac{K}{\delta} + d\log n_i}{n_i}}\right)$$

  i.e. dominated by sample size of the *smallest* group

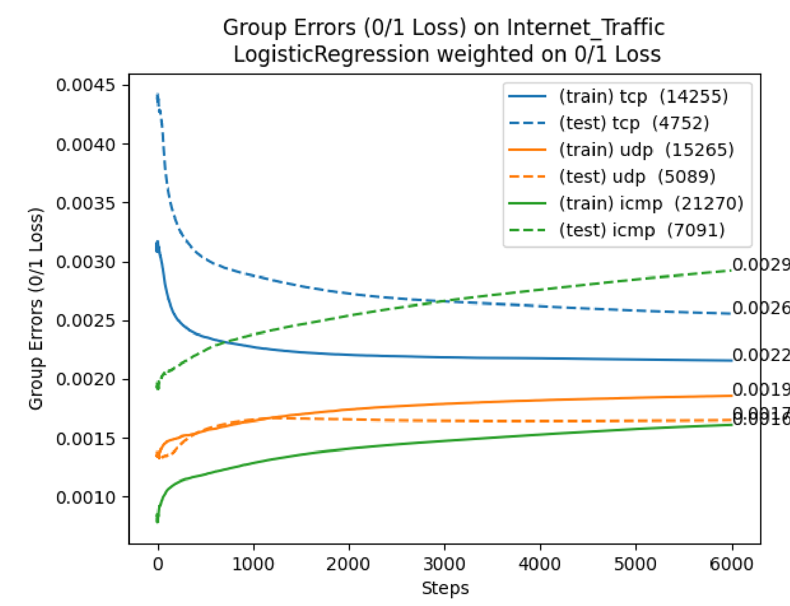
Generalization Experiments

Train vs. Test Performance of MINIMAXFAIR

Bike Dataset



Internet Traffic Dataset



Dataset: Network connection data used to distinguish between 'bad' connections, called intrusions or attacks, and 'good' normal connections.
Label: Connection Legitimacy, *Group*: Protocol Type

Selected References

Natalie Martinez, Martin Bertran, and Guillermo Sapiro (2020)
Minimax Pareto Fairness: A Multi Objective Perspective, *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020.

ProPublica (2020)
COMPAS Recidivism Risk Score Data and Analysis, *Broward County Clerkâs Office, Broward County Sherrif's Office, Florida Department of Corrections, ProPublica*

S. Moro and P. Cortez and P. Rita (2014)
A data-driven approach to predict the success of bank telemarketing, *Decis. Support Syst.*, Volume 52, pgs 22-31.

Sathishkumar V E and Yongyun Cho (2020)
A rule-based model for Seoul Bike sharing demand prediction using weather data, *European Journal of Remote Sensing*, Taylor Francis, pgs 1-18

U. S. Department of Commerce, Bureau of the Census (1990)
Census Of Population And Housing 1990 United States