Adaptive Data Debiasing through Bounded Exploration

Yifan Yang[1*], Yang Liu[2], Parinaz Naghizadeh[1]
[1]: Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio, USA
[2]: Department of Computer Science and Engineering, University of California Santa Cruz, Santa Cruz, California, USA
[*]: Corresponding author: yang.5483@osu.edu

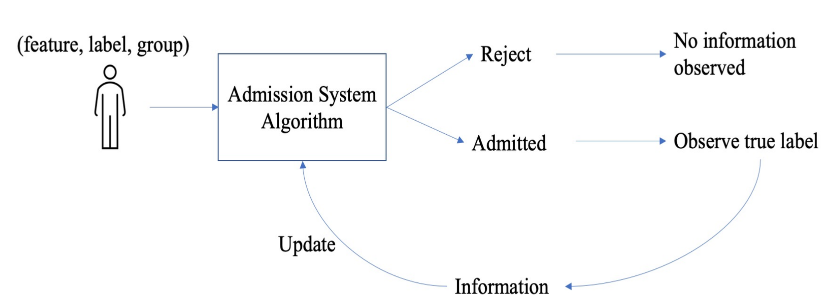
INTRODUCTION

Data-driven decision making algorithm has been adopted widely to help human to make "correct" decisions. In fact, any data-driven algorithm is only as good as the data used to train it. However, bias can exist in most real-life applications due to various reasons. Hence, removing the bias while making accurate and fair decisions becomes important especially in online decision making with partial feedback.

Motivated by this, we propose an algorithm which, while attempting to make accurate (and fair) decisions, also aims to recover unbiased estimates of the underlying distribution of agents interacting with it.

PROBLEM STATEMENT

Agent from current time step (e.g., College applicant) comes into system with observable features $x \in R^n$ (e.g., GPA, GRE, etc.), true label $y \in \{0,1\}$ (e.g., Un/qualified) and group $g \in \{a, b\}$ (e.g., M/F). Once admitted, his/her features will be added into database, which will be used to help decision maker to decide for next time step. The diagram can be represented as follows:



ASSUMPTIONS

- $(x', y, g) = (x', 0/1, a/b)$, where x' is 1-d feature score. If $x \in R^n$, then reduce it from $x \to x'$.
- All incoming agents are i.i.d. from the true distribution $f(\mu_g^y) \sim F$, where f is given from distribution family F with single parameter μ_g^y unknown.
- All existing agents are from initial biased distribution $f(\hat{\mu}_{g,t=0}^y) \sim F$

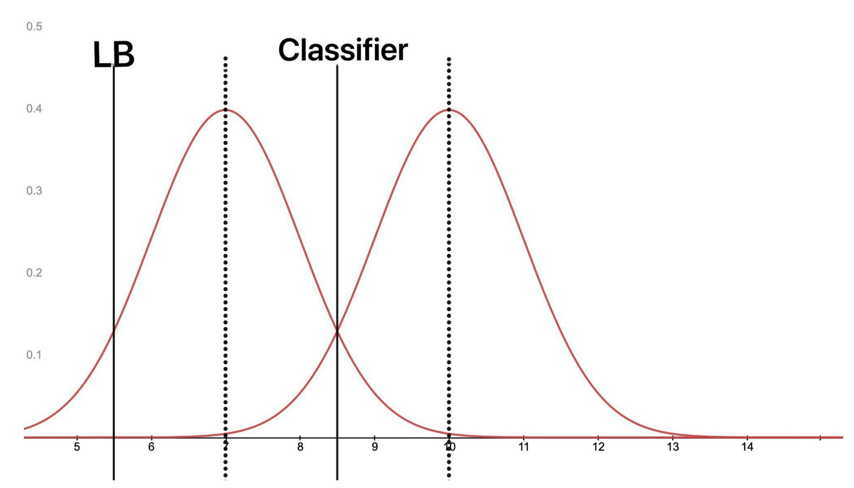
GOAL

- Remove the bias between $\hat{\mu}_{g,t=0}^y$ and μ_g^y
- Obtain true classifier
- Make more correct and fair decisions

LB FOR TRUNCATION

$$LB = \left(\hat{F}^0\right)^{-1}\left(2\hat{F}^0(\hat{\mu}^0) - \hat{F}^0(\theta)\right)$$

Where $\hat{F}^y$, $\left(\hat{F}^y\right)^{-1}$, $\hat{\mu}^y$ are the cdf, inverse cdf and reference value of the estimated distribution $\hat{f}^y$.



DEBIASING ALGORITHM

Step 1: Dimension reduction $R^n \to R$: $x \to x'$:

$$\log \frac{x'}{1-x'} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where $x = [x_1, \ldots, x_n] \subset R^n$, $x' \subset R$.

Step 2: Find label 0 reference position in truncated interval [LB, ∞)
(Similar step for label 1, and the other group)

Step 3: Admit agents (d=1) and collect data:

$$\text{Decision } d = \begin{cases} 1 & x' \geq \theta \\ 1 & \text{LB} \leq x' \leq \theta \text{ with prob } \epsilon \\ 0 & \text{Otherwise} \end{cases}$$

Step 4: Update label 0 distribution estimates:

$$\text{Use new accepted data in} \begin{cases} [\text{LB}, \theta] & \textit{with prob.} 1 \\ [\theta, \infty) & \textit{with prob.} \epsilon \end{cases}$$

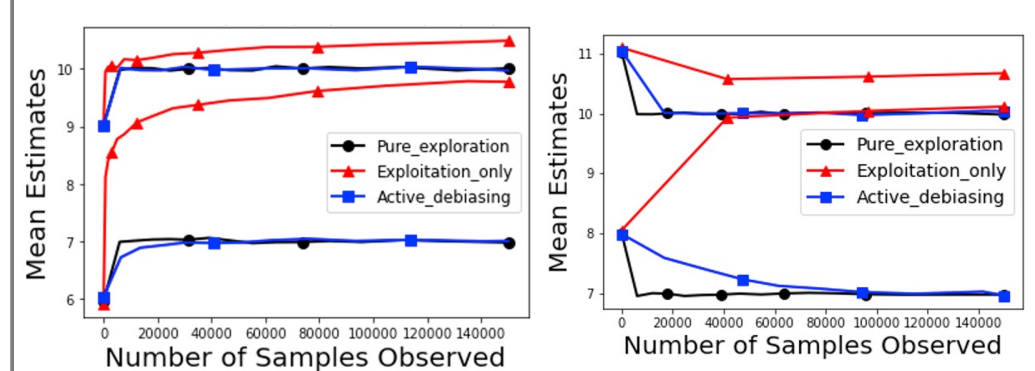
to find the new reference position.

EXPERIMENT

Dataset: Synthetic, Adult and FICO
Comparison: 1: exploitation_only (no explore)
2: pure_exploration (no LB/UB)
3: adaptive (explore with LB)

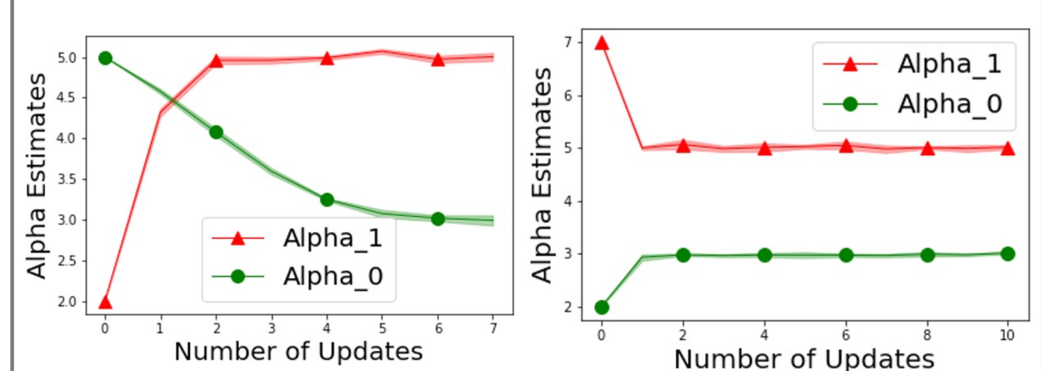
Gaussian on Synthetic

True Distributions:	Initial Biased Distributions:
Label 1: N(10, 1)	Label 1: left: N(9, 1) right: N(11,1)
Label 0: N(7, 1)	Label 0: left: N(6, 1) right: N(8,1)



Beta on Synthetic

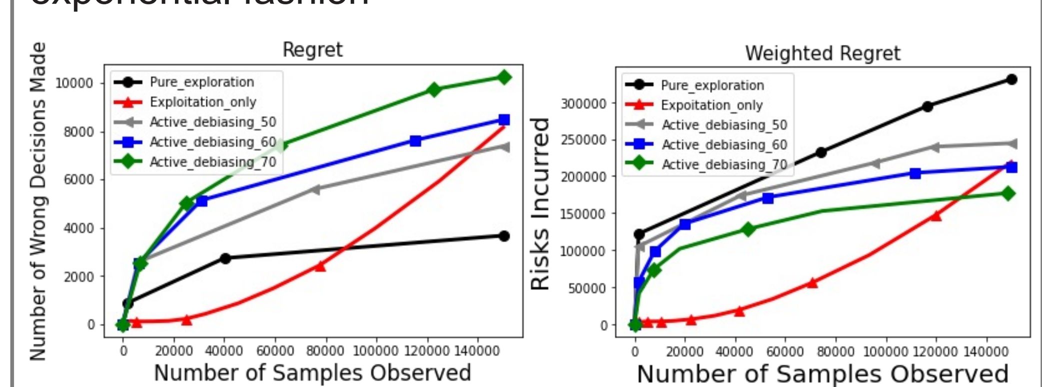
True Distributions:	Initial Biased Distributions:
Label 1: Beta(5, 3)	Label 1: left: Beta(2, 3) right: Beta(7,3)
Label 0: Beta(3, 5)	Label 0: left: Beta(5, 5) right: Beta(2,5)



Regret and Weighted Regret

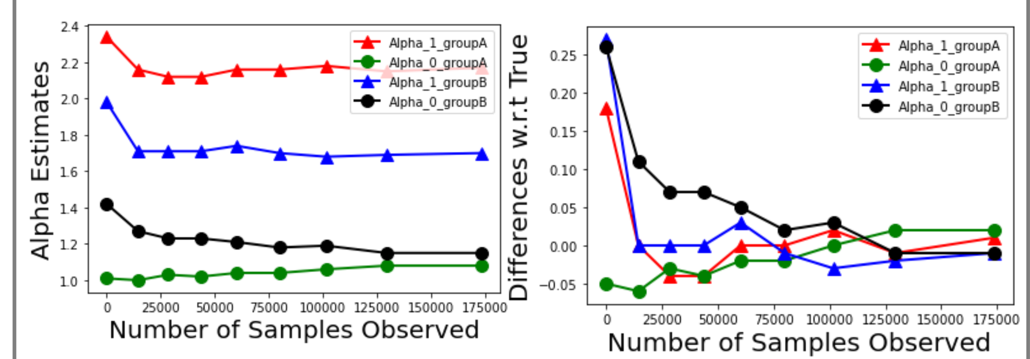
Regret = (FN+FP)_model – (FN+FP)_optimal

Weighted Regret has similar expression in exponential fashion



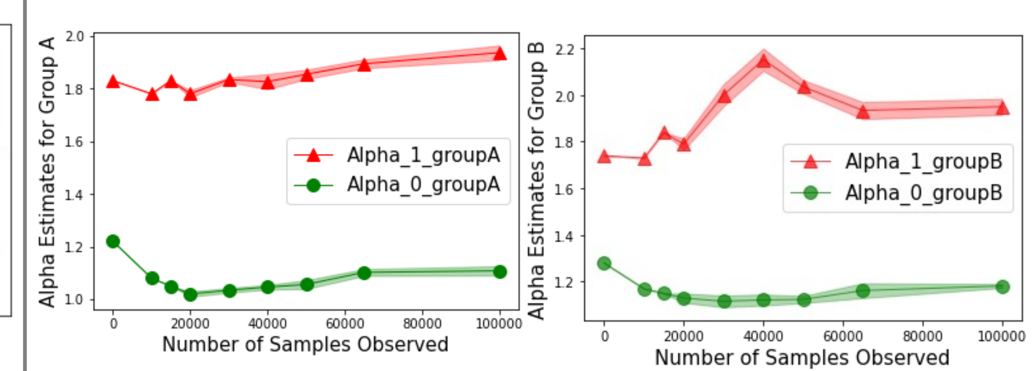
FICO Dataset with EO

True Distributions:	Initial Biased Distributions:
f_a^1 : Beta(2.16, 1.27)	$\hat{f}_{a,t=0}^1$: Beta(2.34, 1.27)
f_a^0 : Beta(1.06, 3.98)	$\hat{f}_{a,t=0}^0$: Beta(1.01, 3.98)
f_b^1 : Beta(1.71, 1.62)	$\hat{f}_{b,t=0}^1$: Beta(1.98, 1.62)
f_b^0 : Beta(1.16, 5.51)	$\hat{f}_{b,t=0}^0$: Beta(1.42, 5.51)



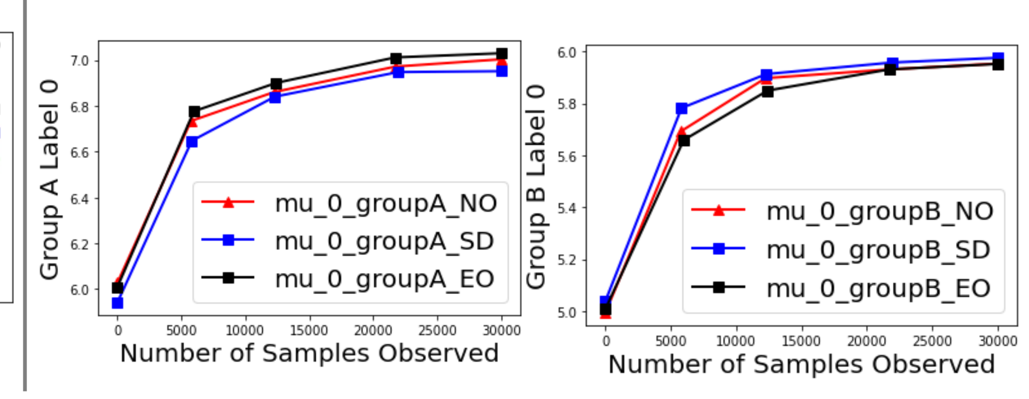
Adult Dataset with EO

True Distributions:	Initial Biased Distributions:
f_a^1 : Beta(1.94, 3.32)	$\hat{f}_{a,t=0}^1$: Beta(1.83, 3.32)
f_a^0 : Beta(1.13, 4.99)	$\hat{f}_{a,t=0}^0$: Beta(1.22, 4.99)
f_b^1 : Beta(1.97, 3.53)	$\hat{f}_{b,t=0}^1$: Beta(1.74, 3.53)
f_b^0 : Beta(1.19, 6.10)	$\hat{f}_{b,t=0}^0$: Beta(1.26, 6.10)



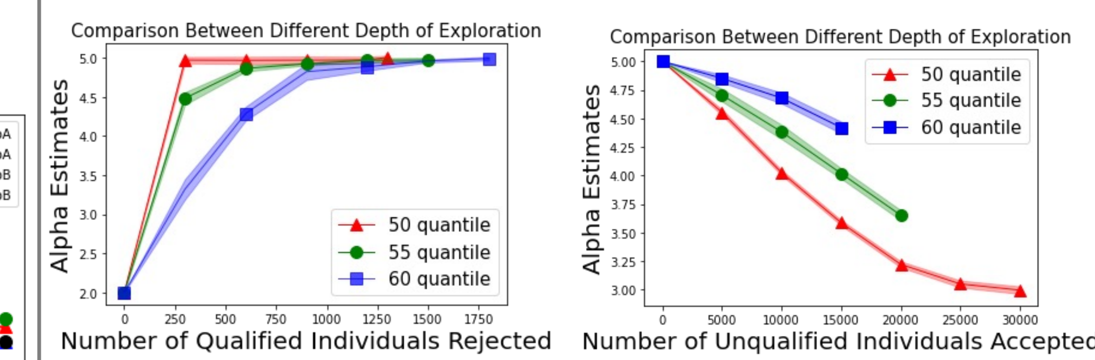
FAIRNESS IMPACT

SD will over-select the majority group (e.g., $\theta_a^F < \theta_a^U$). As an opposite effect, it will under-select the minority group (e.g., $\theta_b^U < \theta_b^F$).



DEPTH OF EXPLORATION IMPACT

One can choose difference reference value of the estimated distribution $\hat{f}^y$. For example, the 50-th quantile corresponds to the median. The experiment is conducted with Beta distribution.



CONCLUSIONS

We contrast our proposed algorithm against two baselines, and show that

- Exploitation-only always leads to overestimates of the underlying distributions.
- Pure_exploration can debias the distribution estimates in the long-run, but it is costly.
- Our algorithm has lower learning regret than exploitation_only, and lower weighted regret than pure_exploration.
- Experiments show that our algorithm can also achieve the same result for asymmetric distribution (e.g., FICO, Adult, Beta Synthetic)
- We analyze the impact of fairness constraints on our algorithm's performance.
- We provide analytical support that our proposed active debiasing algorithm can correct biases in unimodel distribution estimates. We also provide an error bound analysis for our algorithm.

ACKNOWLEDGEMENTS

The authors are grateful for support from Cisco Research, and the NSF program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040800. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, Amazon, or Cisco.