Fairness for Robust Log Loss Classification

Ashkan Rezaei[1], Rizal Fathony[2], Omid Memarrast[1], Brian Ziebart[1]
[1]*Department of Computer Science, University of Illinois at Chicago*
[2]*School of Computer Science, Carnegie Mellon University*

Introduction

- We re-derive a new classifier from the first principles of **distributional robustness** that incorporates **group fairness** criteria into a worst-case logarithmic loss minimization.
- Given population distribution $(\mathbf{X}, A, Y) \sim P$ with a **protected attribute** A and a decision variable $\widehat{Y}$. A classifier $\mathbb{P}$ satisfies group fairness constraints when:
  - **Demographic Parity** (DP): $\mathbb{P}(\widehat{Y}=1|A=a) = \mathbb{P}(\widehat{Y}=1) \quad \forall a \in \{0,1\}$
  - **Equalized Odds** (E.ODD) and **Equalized Opportunity** (E.OPP):

$$\mathbb{P}(\widehat{Y}=1|A=a, Y=y) = \mathbb{P}(\widehat{Y}=1|Y=y), \qquad \begin{array}{l}\text{E.ODD: } \forall a, y \in \{0,1\}.\\ \text{E.OPP: } \forall a \in \{0,1\}, y=1.\end{array}$$

- Our formulation forms a minimax game and produces a parametric exponential family conditional distribution that resembles **truncated logistic regression**.

Robust Log Loss Formulation

- A distributionally robust approach
- Construct predictor robust to worst plausible reality

$$\min_{\mathbb{P}(\hat{y}|\mathbf{x}) \in \Delta} \max_{\mathbb{Q}(\check{y}|\mathbf{x}) \in \Delta \cap \Xi} -\sum_{\mathbf{x}, \hat{y}} \widetilde{P}(\mathbf{x}) \mathbb{Q}(\check{y}|\mathbf{x}) \log \mathbb{P}(\hat{y}|\mathbf{x}) = \max_{\widehat{P}(\check{y}|\mathbf{x}) \in \Xi} H(\widehat{Y}|\mathbf{X})$$

$$\text{, subject to: } \Xi : \left\{ \mathbb{Q} \mid \mathbb{E}_{\widetilde{P}(\mathbf{x}); \mathbb{Q}(\check{y}|\mathbf{x})}[\phi(\mathbf{X}, \check{Y})] = \mathbb{E}_{\widetilde{P}(\mathbf{x},y)}[\phi(\mathbf{X}, Y)] \right\},$$

- Reduces to ***Logistic Regression***: $\mathbb{P}(\hat{y}=1|\mathbf{x}) = e^{\theta^\top \phi(\mathbf{x},1)} / Z_\theta(\mathbf{x})$

Fair Robust Log Loss Formulation

Add fairness constraint to predictor:

$$\min_{\mathbb{P} \in \Delta \cap \Gamma} \max_{\mathbb{Q} \in \Delta \cap \Xi} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \mathbb{Q}(\check{y}|\mathbf{x},a,y)}} \left[-\log \mathbb{P}(\widehat{Y}|\mathbf{X}, A, Y) \right].$$

The sets of decision functions $\mathbb{P}$ satisfying these fairness constraints are *convex* and can be defined using linear constraints:

$$\Gamma : \left\{ \mathbb{P} \mid \frac{1}{p_{\gamma_1}} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \mathbb{P}(\hat{y}|\mathbf{x},a,y)}} [\mathbb{I}(\widehat{Y}=1 \wedge \gamma_1(A,Y))] = \frac{1}{p_{\gamma_0}} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \mathbb{P}(\hat{y}|\mathbf{x},a,y)}} [\mathbb{I}(\widehat{Y}=1 \wedge \gamma_0(A,Y))] \right\},$$

- γ_1 and γ_0 denote some combination of group membership and ground-truth.
- p_{γ_1} and p_{γ_0} denote the empirical frequencies of γ_1 and γ_0: $p_{\gamma_i} = \mathbb{E}_{\widetilde{P}(a,y)}[\gamma_i(A,Y)]$.
- We specify γ_1 and γ_0 for each fairness constraints as:

$$\Gamma_{\text{dp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j); \tag{1}$$

$$\Gamma_{\text{e.opp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j \wedge Y=1); \tag{2}$$

$$\Gamma_{\text{e.odd}} \iff \gamma_j(A,Y) = \begin{bmatrix} \mathbb{I}(A=j \wedge Y=1) \\ \mathbb{I}(A=j \wedge Y=0) \end{bmatrix}. \tag{3}$$

with **Lagrange multipliers** θ for moment matching and λ for fairness constraints, respectively, and n samples in the dataset. The parametric distribution of $\mathbb{P}$ is:

$$\mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y) = \begin{cases} \min\left\{ e^{\theta^\top \phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), \frac{p_{\gamma_1}}{\lambda} \right\} & \text{if } \gamma_1(a,y) \wedge \lambda > 0 \\ \max\left\{ e^{\theta^\top \phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), 1 - \frac{p_{\gamma_0}}{\lambda} \right\} & \text{if } \gamma_0(a,y) \wedge \lambda > 0 \\ \max\left\{ e^{\theta^\top \phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), 1 + \frac{p_{\gamma_1}}{\lambda} \right\} & \text{if } \gamma_1(a,y) \wedge \lambda < 0 \\ \min\left\{ e^{\theta^\top \phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), -\frac{p_{\gamma_0}}{\lambda} \right\} & \text{if } \gamma_0(a,y) \wedge \lambda < 0 \\ e^{\theta^\top \phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}) & \text{otherwise,} \end{cases}$$

where $Z_\theta(\mathbf{x}) = e^{\theta^\top \phi(\mathbf{x},1)} + e^{\theta^\top \phi(\mathbf{x},0)}$ is the normalization constant.

- Jointly optimize λ and θ.
- Given θ we find optimal λ^* (the threshold) in $O(n \log n)$ over n-sample batch.
- Given λ^* the objective is convex w.r.t $\theta \rightarrow$ employ *batch* gradient decent.

Analysis

- Provides a **monotonic** and **parametric** transformation of probabilities.

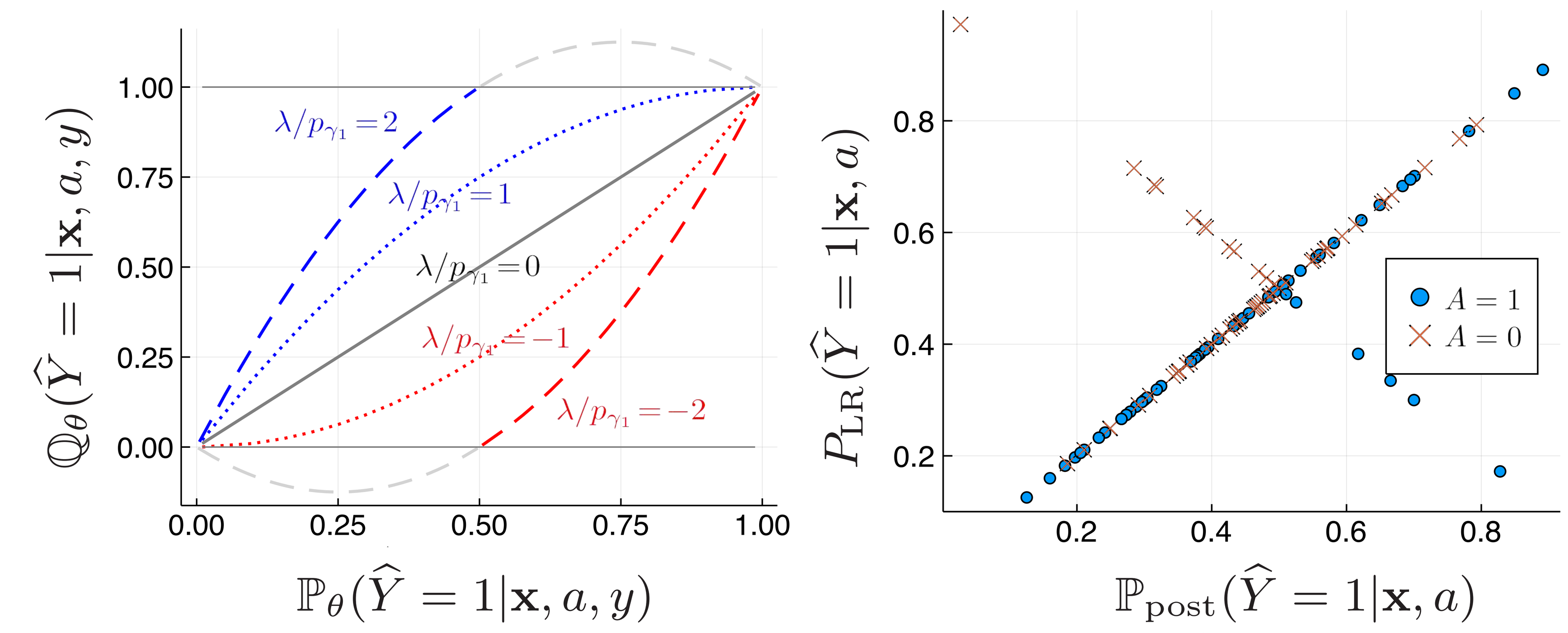


Figure 1. Contrast the relationship between predictor ($\mathbb{P}$) and approximator's ($\mathbb{Q}$) parametric distributions in our method (left) and the post-processing (Hardt et al. 2016) transformation of logistic regression prediction (right).

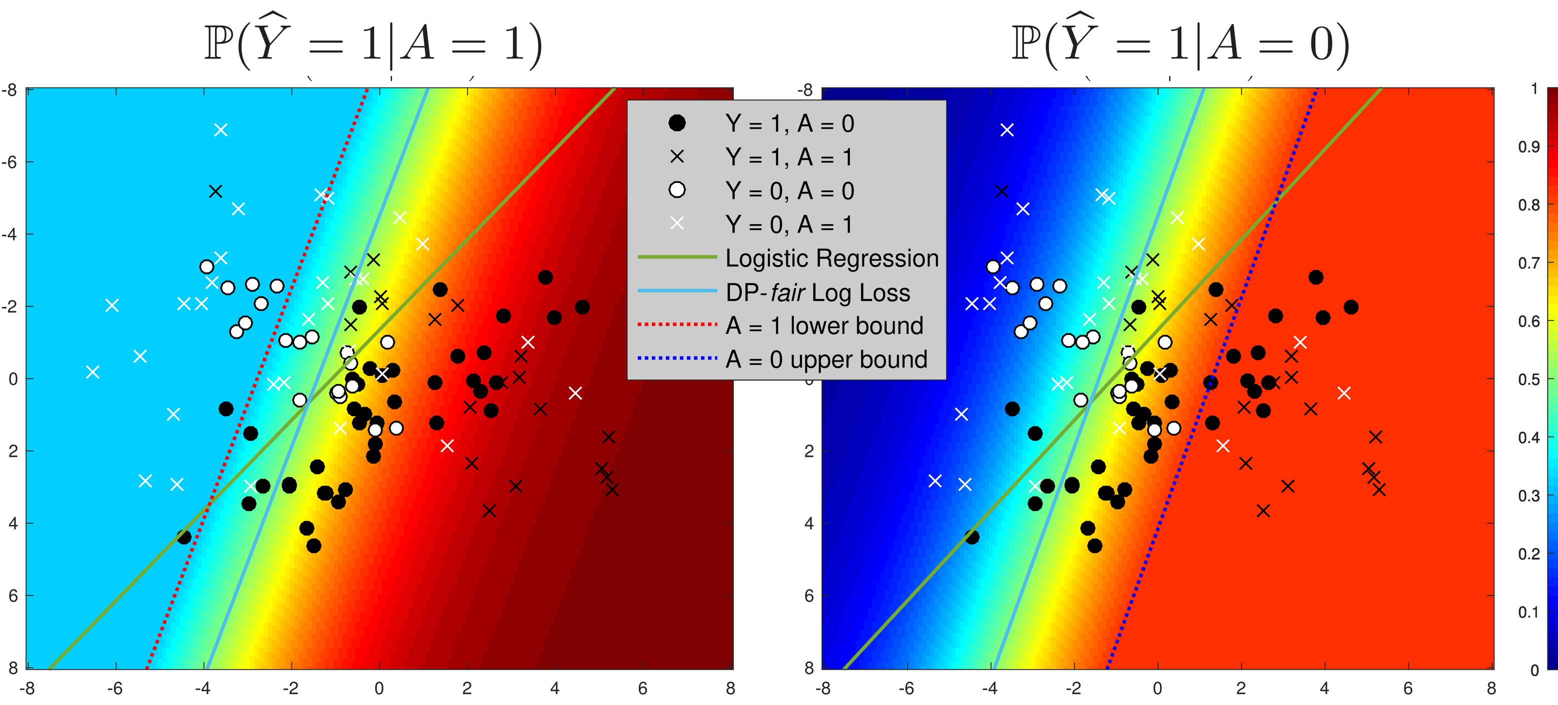
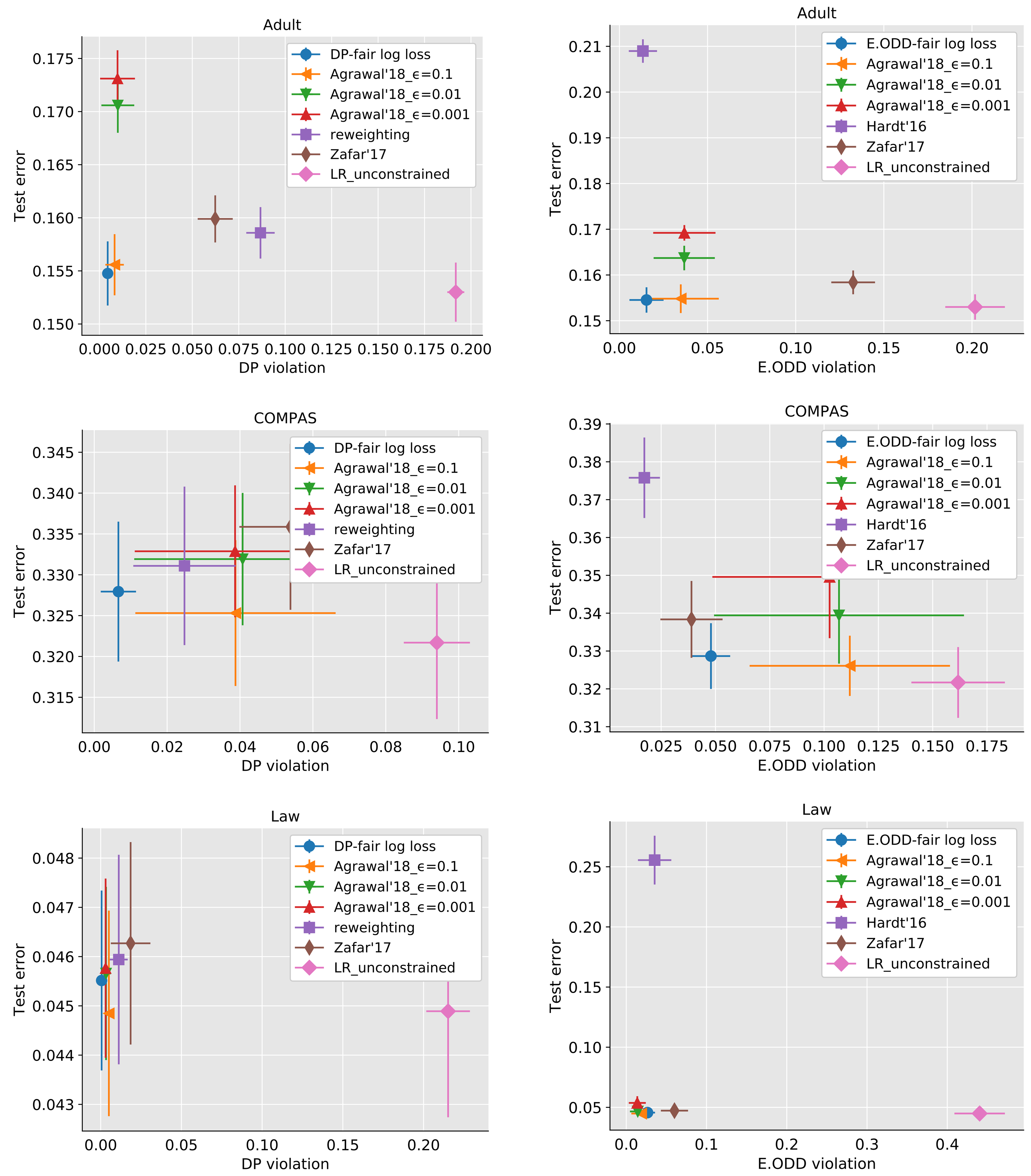


Figure 2. Experimental results on a synthetic dataset with: a heatmap indicating the predictive probabilities of our approach, along with **decision** and **threshold boundaries**; and the *unfair* logistic regression decision boundary.

Experiments



- Our method reside in *Pareto optimal* set: none of the other baselines are significantly better than our method on both error and fairness violation.
- Order of magnitude improvement in running time compared to reduction-based approach methods of Agrawal et al. 2018 and covariance-proxy approach of Zafar et al. 2017.

Acknowledgment: This work was supported, in part, by the National Science Foundation under Grant No. 1652530 and by the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program.