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 derive a new classifier from the first principles of **distributional robustness** that finds a fair predictor against a robust adversary (i.e., *worst-case approximation of the training data).*

• Our **in-process** minimax game formulation produces a parametric exponential family conditional distribution that resembles **truncated logistic regression**.

• Our predictor is **uniquely defined**, **robust against label noise**, **asymptotically consistent**, and **respectful of group fairness requirements.**

Robust Log Loss Formulation

• Construct predictor robust to worst plausible training data labels:

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

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

• Reduces to ***Logistic Regression***: $\mathbb{P}(\widehat{y}=1|\mathbf{x}) = e^{\theta^{\mathrm{T}}\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x})$.

Fair Robust Log Loss Formulation

Enforce fairness constraint (Γ) on robust predictor:

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

with **protected attribute** A and decision variable $\widehat{Y}$.

Group fairness constraints are convex constraint sets (Agarwal et al. 2018):

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

where p_{γ_j} denote the empirical frequencies of γ_j: $p_{\gamma_j} = \mathbb{E}_{\widetilde{P}(a,y)}[\gamma_j(A,Y)]$

Demographic Parity: $\mathbb{P}(\widehat{Y}=1|A=j) = \mathbb{P}(\widehat{Y}=1)$

$$\Gamma_{\text{dp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j) \qquad \forall j\in\{0,1\}$$

Equalized Opportunity: $\mathbb{P}(\widehat{Y}=1|A=j,Y=1) = \mathbb{P}(\widehat{Y}=1|Y=1)$

$$\Gamma_{\text{e.opp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j\wedge Y=1) \qquad \forall j\in\{0,1\}$$

Equalized Odd: $\mathbb{P}(\widehat{Y}=1|A=j,Y=y) = \mathbb{P}(\widehat{Y}=1|Y=y)$

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

• With **Lagrange multipliers** θ for moment matching and λ for fairness constraints, respectively, The parametric distribution of $\mathbb{P}$ is:

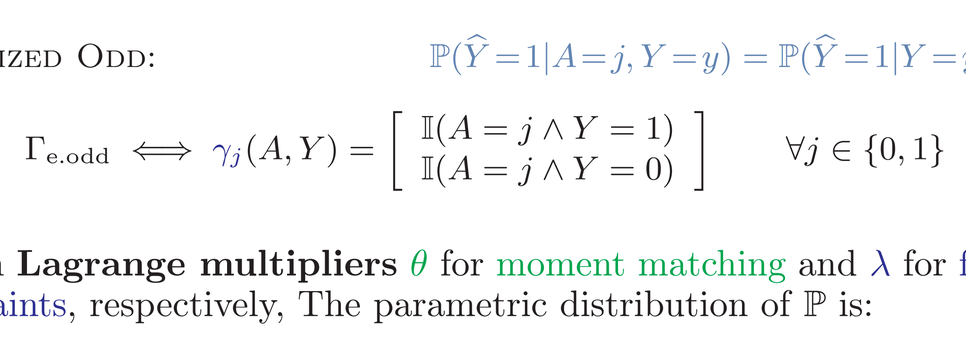
$$\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \begin{cases}\min\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), \frac{p_{\gamma_1}}{\lambda}\right\} & \text{if } \gamma_1(a,y)\wedge\lambda>0\\ \max\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}), 1-\frac{p_{\gamma_0}}{\lambda}\right\}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

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

• Given θ we find optimal threshold λ^* in $O(n\log n)$ over n-sample batch.

• The objective is convex w.r.t $\theta\rightarrow$ employ *batch* gradient descent.

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



Contrast the relationship between ($\mathbb{P}$) and adversary ($\mathbb{Q}$) in our method (left) and the post-processing (Hardt et al. 2016) transformation of logistic regression prediction (right).

Making Decisions with Label-based Fairness

• For E.ODD and E.OPP we have $\mathbb{P}(\widehat{y}|\mathbf{x},a,y)$ but we need $\mathbb{P}(\widehat{y}|\mathbf{x},a)$ at test time.

• We use our robust estimation of true distribution P and get a fixed-point estimate by marginal probability rule.

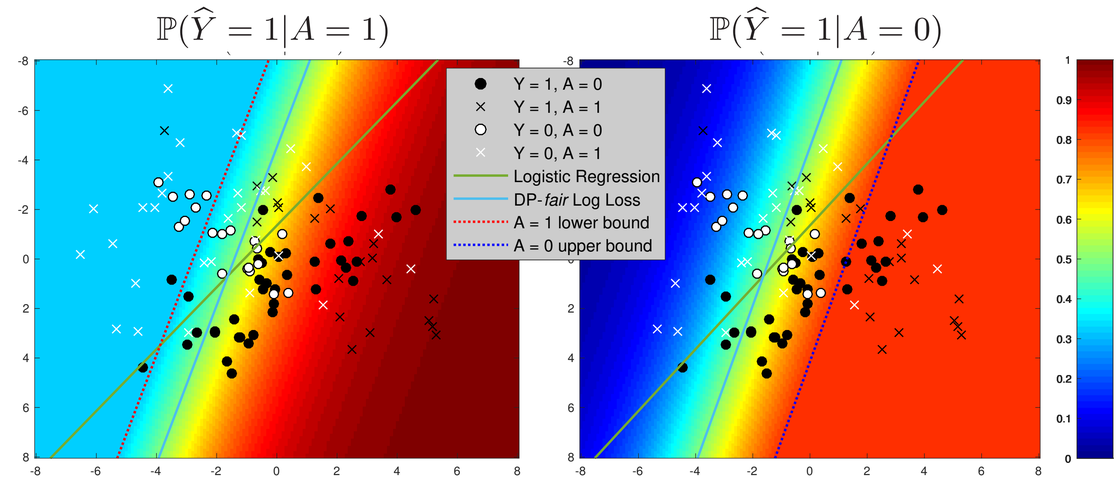
$$\mathbb{Q}(\widehat{y}|\mathbf{x},a) \approx \mathbb{Q}(\widehat{y}|\mathbf{x},a,y=1)\mathbb{Q}(\widehat{y}=1|\mathbf{x},a) + \mathbb{Q}(\widehat{y}|\mathbf{x},a,y=0)\mathbb{Q}(\widehat{y}=0|\mathbf{x},a)$$

• Use $\mathbb{Q}(\widehat{y}|\mathbf{x},a)$ in place of true distribution to marginalize:

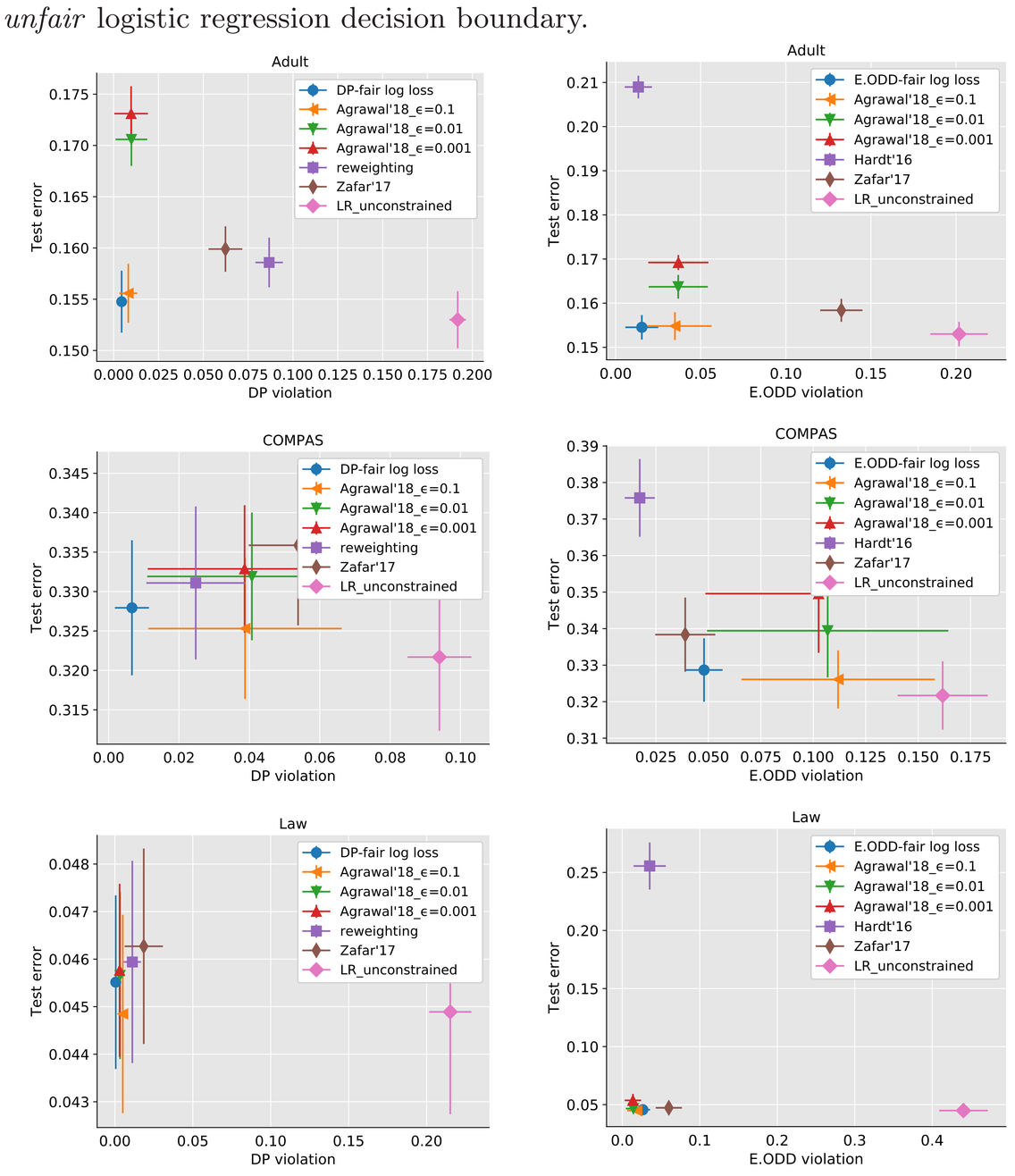
$$\mathbb{P}(\widehat{y}|\mathbf{x},a) = \mathbb{P}(\widehat{y}|\mathbf{x},a,y=1)\mathbb{Q}(\widehat{y}=1|\mathbf{x},a) + \mathbb{P}(\widehat{y}|\mathbf{x},a,y=0)\mathbb{Q}(\widehat{y}=0|\mathbf{x},a)$$

Theorem. *For fairness constraints that depend on the true label, our inference procedure produces the marginal predicting distribution $\mathbb{P}$ of the fair predictor distribution with the closest KL-divergence to $P(\mathbf{x},a,y)$ in the limit.*

Experiments



• Synthetic dataset results with heatmap indicating the predictive probabilities of our approach, along with **decision** and **threshold boundaries**; and the *unfair* logistic regression decision boundary.



• Our method resides in the **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 *cost-sensitive reduction* approach of Agarwal et al. 2018 and *covariance constraint* 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.

(*) These two authors contributed equally