Robust Fairness Under Covariate Shift

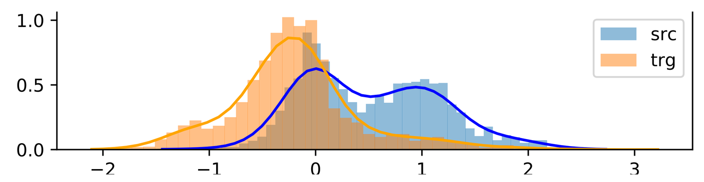
Ashkan Rezaei[1], Anqi Liu[2], Omid Memarrast[1], Brian Ziebart[1]
[1] University of Illinois at Chicago,
[2] Caltech

THE UNIVERSITY OF ILLINOIS AT CHICAGO

Caltech

Introduction

- We seek fairness for classification under **non-IID** assumption.
- **Covariate Shift** assumption: the inputs/covariates change $P_{\mathbf{src}}(\mathbf{x}) \neq P_{\mathbf{trg}}(\mathbf{x})$ while the conditional label distribution $P(y|\mathbf{x})$ remains the same.



- We seek fair decisions (true/false positive rate parity) on **target data** with **unknown labels**.
- We take **distributionally robust** approach to obtains a predictor that is **robust** against an adversary which approximates **worst-case target performance** penalized by fairness cost while **matching source data** on feature statistics.

Robust Log Loss Under Covariate Shift [Liu and Ziebart (2014)]

- Construct predictor robust to worst plausible training data labels:

$$\min_{\mathbb{P}(y|\mathbf{x})\in\Delta} \max_{\mathbb{Q}(y|\mathbf{x})\in\Delta\cap\Xi} \mathbb{E}_{P_{\text{trg}}(\mathbf{x})\mathbb{Q}(y|\mathbf{x})}[-\log \mathbb{P}(Y|\mathbf{X})]$$
$$= \max_{\mathbb{P}(y|\mathbf{x})\in\Delta\cap\Xi} H_{P_{\text{trg}}(\mathbf{x})\mathbb{P}(y|\mathbf{x})}(Y|\mathbf{X}),$$

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

- Reduces to following parametric form:

$$\mathbb{P}_\theta(y|\mathbf{x}) = e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y)} \Big/ \sum_{y'\in\mathcal{Y}} e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y')}$$

Fairness Under Covariate Shift

- True positive rate parity (equalized opportuity):

$$P(\hat{Y}=1|A=1, Y=1) = P(\hat{Y}=1|A=0, Y=1).$$

A: **protected attribute**, $\hat{Y}$: decision variable and Y: true label.

- Most IID methods, infer fairness from training data where Y is **observed** $\rightarrow$ Fairness is a linear constraint on $\hat{Y}$.
- We seek to ensure fairness on **target** data, with **unknown true label** $\rightarrow$ random variable $Y \sim$ worst-case estimate $\mathbb{Q}$
- Fairness becomes a bi-linear constraint on target data $\rightarrow$ we enforce by **penalty** term.

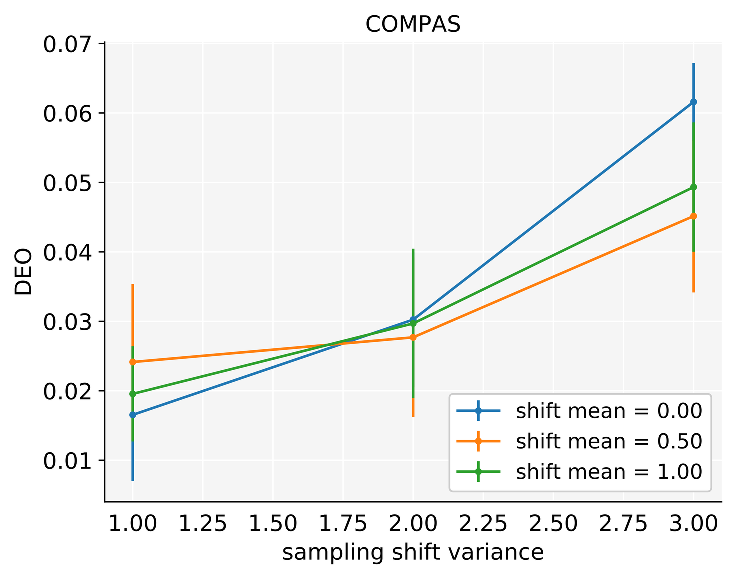


Figure: The DEO progression of fairLR [Rezaei et al. 2020], with increasing distribution shift

Our Model

Our fair predictor $\mathbb{P}$ minimizes the **worst-case expected log loss** with an **μ-weighted expected fairness penalty** on **target**, approximated by adversary $\mathbb{Q}$ constrained to **match source distribution statistics** (Ξ) and **group marginals on target** (Γ):

$$\min_{\mathbb{P}\in\Delta} \max_{\mathbb{Q}\in\Delta\cap\Xi\cap\Gamma} \mathbb{E}_{P_{\text{trg}}(\mathbf{x},a)\mathbb{Q}(y|\mathbf{x},a)}[-\log \mathbb{P}(Y|\mathbf{x},a)] + \mu\, \mathbb{E}_{P_{\text{trg}}(\mathbf{x},a)\mathbb{Q}(y'|\mathbf{x},a)\mathbb{P}(y|\mathbf{x},a)}[f(A,Y',Y)] \quad (1)$$

such that:

$$\Xi(\mathbb{Q}) : \mathbb{E}_{\substack{P_{\text{src}}(\mathbf{x},a)\\ \mathbb{Q}(y|\mathbf{x},a)}}[\phi(\mathbf{X},Y)] = \mathbb{E}_{P_{\text{src}}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)] \text{ and}$$
$$\forall k \in \{0,1\},$$
$$\Gamma(\mathbb{Q}) : \mathbb{E}_{\substack{P_{\text{trg}}(\mathbf{x},a)\\ \mathbb{Q}(y|\mathbf{x},a)}}[g_k(A,Y)] = \underbrace{\mathbb{E}_{\substack{P_{\text{trg}}(\mathbf{x},a)\\ \hat{P}_{\text{trg}}(y|\mathbf{x},a)}}[g_k(A,Y)]}_{\tilde{g}_k},$$

where:

- ϕ is the feature function, e. g: $\phi(\mathbf{x},y) = [x_1 y, x_2 y, \dots x_m y]^\top$.
- μ is the fairness penalty weight.
- Ξ is feature-matching on **source**, Γ is group marginal matching on **target**
- $g_k(.,.)$ is a group k selector function, i.e. for equalized opportunity: $g_k(A,Y) = \mathbb{I}(A{=}k \wedge Y{=}1)$
- $\tilde{g}_k$ is the group k density on target, *estimated offline*
- $f(.,.,.)$ is a weighting function of the mean score difference between the two groups:

$$f(A,Y,\hat{Y}) = \begin{cases} \frac{1}{\tilde{g}_1} & \text{if } g_1(A,Y) \wedge \mathbb{I}(\hat{Y}{=}1) \\ -\frac{1}{\tilde{g}_0} & \text{if } g_0(A,Y) \wedge \mathbb{I}(\hat{Y}{=}1) \\ 0 & \text{otherwise.} \end{cases}$$

Theorem. *The predictor $\mathbb{P}$ in our model (1) for a given fairness penalty weight μ, can be obtained by solving:*

$$\log \frac{1-\mathbb{P}(y|\mathbf{x},a)}{\mathbb{P}(y|\mathbf{x},a)} + \mu\, \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)}[f(a,y,Y')] + \frac{P_{src}(\mathbf{x},a)}{P_{trg}(\mathbf{x},a)}\theta^{\text{T}}(\phi(\mathbf{x},y=1) - \phi(\mathbf{x},y=0)) + \sum_{k\in 0,1}\lambda_k g_k(a,y) = 0,$$

where :

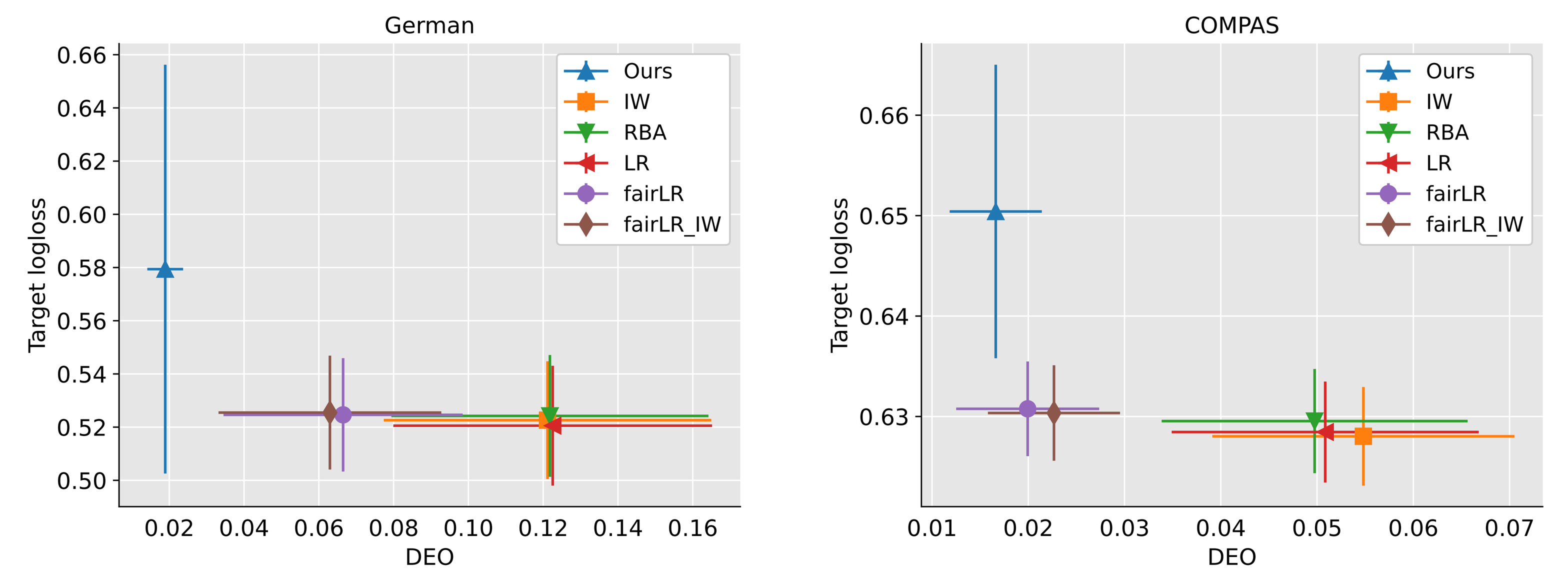
- θ and λ are the dual Lagrange multipliers for Ξ and Γ constraints respectively.
- Given the solution $\mathbb{P}^*$, the $\mathbb{Q}$ in equilibrium is:

$$\mathbb{Q}(y|\mathbf{x},a) = \frac{\mathbb{P}^*(y|\mathbf{x},a)}{1 - \mu f(a,y,y) + \mu f(a,y,y)\mathbb{P}^{*2}(y|\mathbf{x},a)}$$

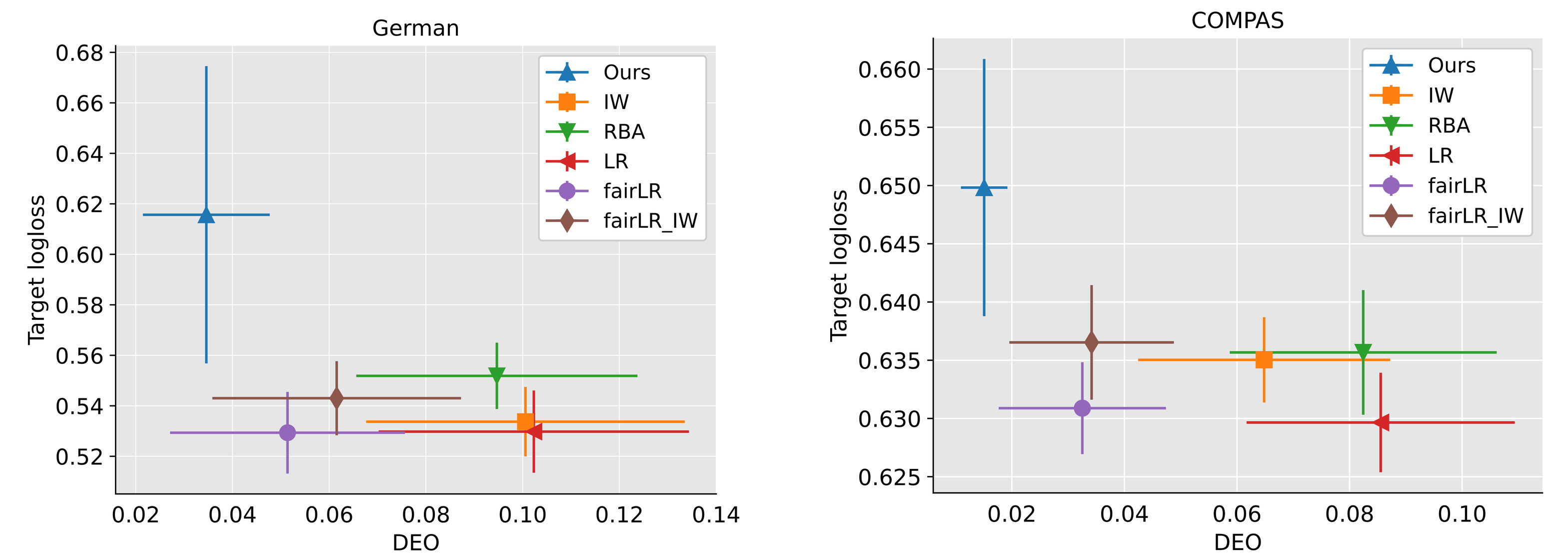
where $0 \leq \mathbb{Q}(y|x,a) \leq 1$.

- We employ a **batch gradient decent** to obtain θ^* and λ^*.
- We **binary-search** for optimal weight μ that makes **expected fairness cost closest to zero**. Assuming sufficient expressive feature constraints, $\mathbb{Q}$ remains monotone in relatively small intervals.

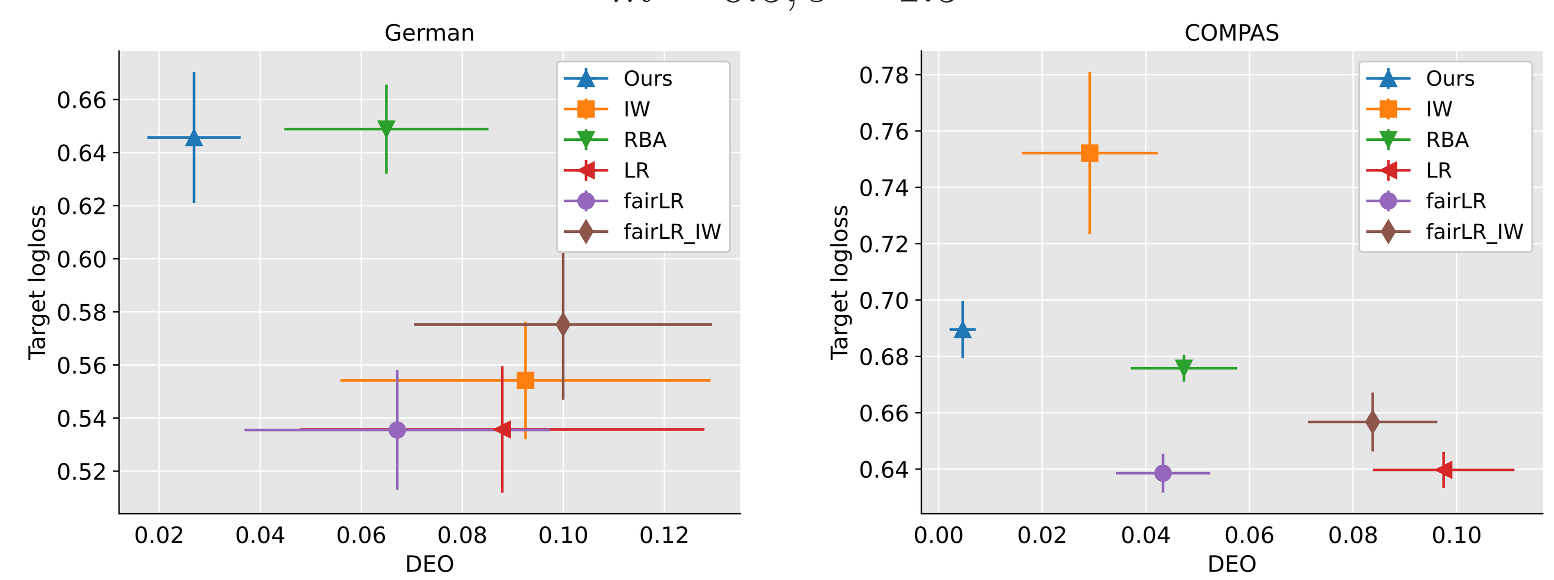
Experiments



$m = 0, s = 1$ (IID)



$m = 0.5, s = 1.5$



$m = 1, s = 2$

- We create covariate shift by biased sampling on first principal component $\mathcal{C}$ of the features, according to a shifted Gaussian $D_{\text{src}}(\mu(\mathcal{C}) + m, \frac{\sigma(\mathcal{C})}{s})$.
- As the shift increases our method fairness violation stays low, with logloss trade-off compared to other methods.