Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification (NeurIPS 2019)

Sungeun Lee

September 28, 2025

Seoul National University

Contents

Introduction

Notations

- $X \in \mathbb{R}^d$: d-dimensional feature vector,
- $S \in \{0, 1\}$: Binary Sensitive Attribute,
- $Y \in \{0, 1\}$: Binary Label,
- $g : \mathbb{R}^d \times \{0, 1\} \to \{0, 1\}$: Classifier (a measurable function),
- $\eta(x, s) := \mathbb{P}(Y = 1 \mid X = x, S = s)$: Regression Function
- $\mathcal{R}(g) := \mathbb{P}(g(X, S) \neq Y)$: Risk Function.

Problem Motivation

1. Machine learning is widely deployed in society, but models often produce **unfair outcomes** when predictions correlate with **sensitive attributes** (e.g. gender, race).
2. The paper focuses on **Equal Opportunity(EO) Fairness**, which requires equal True Positive Rates(TPR) across sensitive groups.

▶ **Equal Opportunity**

A classifier $g(x, s) \in \{0, 1\}$ is called fair if

$$P\big(g(X, S) = 1 \,\big|\, S = 1, Y = 1\big) = P\big(g(X, S) = 1 \,\big|\, S = 0, Y = 1\big).$$

The set of all fair classifiers is denoted by $\mathcal{F}(\mathbb{P})$.

What We Show?

Among all EO-fair classifiers $\mathcal{F}(\mathbb{P})$, which classifier minimizes the risk?
Answering the question above is equivalent to solving the following problem:

$$\min_{g \in \mathcal{F}(\mathbb{P})} \mathcal{R}(g), \qquad \mathcal{R}(g) := \mathbb{P}\big(g(X, S) \neq Y\big).$$

Equivalently,

$$\min_{g} \mathcal{R}(g) \text{ s.t. } \mathbb{P}(g{=}1 \mid Y{=}1, S{=}1) = \mathbb{P}(g{=}1 \mid Y{=}1, S{=}0).$$

Our paper shows that the EO–optimal classifier takes the following form:

$$g^*(x,1) = \mathbf{1}\Big\{1 \le \eta(x,1)\Big(2 - \tfrac{\theta^*}{\mathbb{P}(Y=1,S=1)}\Big)\Big\}, \quad g^*(x,0) = \mathbf{1}\Big\{1 \le \eta(x,0)\Big(2 + \tfrac{\theta^*}{\mathbb{P}(Y=1,S=0)}\Big)\Big\}.$$

where $\theta^* \in \mathbb{R}$ is chosen to equalize the two groups' TPRs and satisfies $|\theta^*| \le 2$.

What We Show? (continued)

We do not know the true distribution $\mathbb{P}$ nor the oracle shift θ^*. **Our approach** estimates them *empirically* (via a plug-in scheme using labeled data for $\hat{\eta}$ and unlabeled data for $\hat{\theta}$) and then proves **consistency**:

$$\underbrace{\mathbb{E}[\Delta(\hat{g}, \mathbb{P})] \to 0}_{\text{asymptotically fair}} \quad \text{and} \quad \underbrace{\mathbb{E}[\mathcal{R}(\hat{g})] \to \mathcal{R}(g^*)}_{\text{asymptotically optimal}},$$

where $\Delta(g, \mathbb{P}) := \Big| \mathbb{P}(g(X, S) = 1 \mid S = 1, Y = 1) - \mathbb{P}(g(X, S) = 1 \mid S = 0, Y = 1) \Big|$.
Here, $\Delta(g, \mathbb{P})$ is what we call **Unfairness** in this paper.

Methods

Set Up

In this section, we explain ***why the EO-optimal classifier takes the form*** introduced earlier,

$$g^*(x,1) = \mathbf{1}\Big\{1 \le \eta(x,1)\Big(2 - \tfrac{\theta^*}{\mathbb{P}(Y=1,S=1)}\Big)\Big\}, \quad g^*(x,0) = \mathbf{1}\Big\{1 \le \eta(x,0)\Big(2 + \tfrac{\theta^*}{\mathbb{P}(Y=1,S=0)}\Big)\Big\}.$$

how to estimate the unknown quantities in that rule (the distribution $\mathbb{P}$ and the shift θ^*), and ***why the resulting plug-in classifier is consistent***.

▶ **Assumption**

We assume that we have at our disposal two datasets, labeled $\mathcal{D}_n$ and unlabeled $\mathcal{D}_N$:

$$\mathcal{D}_n = \{(X_i, S_i, Y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathbb{P}, \qquad \mathcal{D}_N = \{(X_i, S_i)\}_{i=n+1}^{n+N} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X,S}.$$

Notation for group counts in $\mathcal{D}_N$:

$$N_s := \sum_{i=n+1}^{n+N} \mathbf{1}\{S_i = s\}, \qquad N := \sum_{s\in\{0,1\}} N_s \quad (s \in \{0,1\}).$$

Optimal EO Classifier Proof Sketch (1/3)

Step 1. Wrap EO into a Lagrangian. Start from

$$\min_g \mathcal{R}(g) \quad \text{s.t.} \quad \mathbb{P}(g{=}1 \mid Y{=}1, S{=}1) = \mathbb{P}(g{=}1 \mid Y{=}1, S{=}0).$$

Introduce a multiplier λ and consider the saddle problem $\min_g \max_\lambda \mathcal{L}(g, \lambda)$. By weak duality, solving $\max_\lambda \min_g \mathcal{L}(g, \lambda)$ is enough to identify the optimal form.

$$\mathcal{L}(g, \lambda) = \mathcal{R}(g) + \lambda\Big(\mathbb{P}(g{=}1 \mid Y{=}1, S{=}1) - \mathbb{P}(g{=}1 \mid Y{=}1, S{=}0)\Big).$$

Step 2. Linearize in g. Write both the risk and the EO term as expectations that are *linear in* g:

$$\mathcal{R}(g) = \text{const} - \sum_{s \in \{0,1\}} \mathbb{P}(S{=}s)\, \mathbb{E}_{X \mid S=s}\big[g(X, s)\,(2\eta(X, s) - 1)\big],$$

and the EO difference as $\mathbb{E}_{X|S=s}[\,\eta(X, s)\, g(X, s)\,]/\mathbb{P}(Y{=}1 \mid S = s)$. This lets us minimize $\mathcal{L}(g, \lambda)$ *pointwise* in (x, s).

Optimal EO Classifier Proof Sketch (2/3)

Step 3. Pointwise minimization $\Rightarrow$ threshold rule.
For each (x, s), choose $g(x, s) \in \{0, 1\}$ that minimizes the local linear expression. This yields, for any fixed λ,

$$g_\lambda(x,1) = \mathbf{1}\Big\{1 \le \eta(x,1)\Big(2 - \tfrac{\lambda}{\mathbb{P}(Y=1,S=1)}\Big)\Big\}, \quad g_\lambda(x,0) = \mathbf{1}\Big\{1 \le \eta(x,0)\Big(2 + \tfrac{\lambda}{\mathbb{P}(Y=1,S=0)}\Big)\Big\}.$$

So the solution must be a *thresholded Bayes regressor with group-dependent shift*.

For fixed λ, $\mathcal{L}(g, \lambda)$ is linear in g. The pointwise coefficient of $g(x, s)$ equals

$$\underbrace{-\big(2\eta(x,s) - 1\big)\,\mathbb{P}(S{=}s)}_{\text{from risk}} + \underbrace{\lambda \cdot \frac{\eta(x,1)}{\mathbb{P}(Y{=}1 \mid S{=}1)}\,\mathbf{1}\{s{=}1\} - \lambda \cdot \frac{\eta(x,0)}{\mathbb{P}(Y{=}1 \mid S{=}0)}\,\mathbf{1}\{s{=}0\}}_{\text{from EO term}}.$$

Choose $g(x, s) = 1$ iff this coefficient ≤ 0. For $s{=}1$ it reduces to $1 \le \eta(x,1)\big(2 - \lambda/\mathbb{P}(Y{=}1, S{=}1)\big)$; for $s{=}0$ to $1 \le \eta(x,0)\big(2 + \lambda/\mathbb{P}(Y{=}1, S{=}0)\big)$.

Optimal EO Classifier Proof Sketch (3/3)

Step 4. Pick $\lambda = \theta^*$ to satisfy EO.

Choose θ^* so that the two TPRs match:

$$\frac{\mathbb{E}_{X\mid S=1}\big[\eta(X,1)\, g_{\theta^*}(X,1)\big]}{\mathbb{P}(Y{=}1 \mid S{=}1)} = \frac{\mathbb{E}_{X\mid S=0}\big[\eta(X,0)\, g_{\theta^*}(X,0)\big]}{\mathbb{P}(Y{=}1 \mid S{=}0)}.$$

Under mild continuity, there exists a unique θ^* that equalizes the two TPRs.

Define $\phi(\lambda) := \mathrm{TPR}_1(g_\lambda) - \mathrm{TPR}_0(g_\lambda)$. With no mass at the threshold, ϕ is continuous and strictly monotone in λ, so by the intermediate value theorem there is a unique root.

Step 5. Conclude optimality.

At (g_{θ^*}, θ^*) we satisfy EO and attain the dual optimum. Weak duality then implies g_{θ^*} *minimizes risk* among all EO-fair classifiers.

Step 6. Range of θ^*.

We show

$$2 - \frac{\theta^*}{\mathbb{P}(Y=1, S=1)} > 0 \quad \text{and} \quad 2 + \frac{\theta^*}{\mathbb{P}(Y=1, S=0)} > 0,$$

hence $-2\,\mathbb{P}(Y{=}1, S{=}0) < \theta^* < 2\,\mathbb{P}(Y{=}1, S{=}1)$ and in particular $|\theta^*| \leq 2$.

Proposed Plug-in Procedure (1/2)

Regression estimator
An estimator $\hat{\eta}$ of $\eta(x, s) := \mathbb{P}(Y = 1 \mid X = x, S = s)$ is constructed from the labeled sample $\mathcal{D}_n$ and is independent of the unlabeled sample $\mathcal{D}_N$ (e.g., by sample splitting).

Empirical Distributions $\mathcal{D}_N$
For $s \in \{0, 1\}$, Define

$$\hat{\mathbb{P}}_{X|S=s} = \frac{1}{|\{(X,S)\in\mathcal{D}_N : S=s\}|} \sum_{\{(X,S)\in\mathcal{D}_N : S=s\}} \delta_X , \quad \hat{\mathbb{P}}_S = \frac{1}{N} \sum_{\{(X,S)\in\mathcal{D}_N\}} \delta_S$$

where δ_z denotes the Dirac point mass at z (i.e., the measure that assigns probability 1 to the singleton $\{z\}$ and 0 elsewhere).

Proposed Plug-in Procedure (2/2)

From the optimal-form family

$$g^*(x,1) = \mathbf{1}\Big\{ 1 \le \eta(x,1)\Big(2 - \tfrac{\theta^*}{\mathbb{P}(Y=1,S=1)}\Big)\Big\}, \quad g^*(x,0) = \mathbf{1}\Big\{ 1 \le \eta(x,0)\Big(2 + \tfrac{\theta^*}{\mathbb{P}(Y=1,S=0)}\Big)\Big\}.$$

the unknowns are the joint terms $\mathbb{P}(Y=1, S=s)$ and the regression η.

Use the empirical distributions from $\mathcal{D}_N$.
With the unlabeled sample and $\hat{\eta}$ (trained on $\mathcal{D}_n$), define

$$\widehat{\mathbb{E}}_{X|S=s}[f(X)] := \frac{1}{N_s} \sum_{i=n+1}^{n+N} f(X_i)\,\mathbf{1}\{S_i = s\}, \qquad \widehat{\mathbb{P}}_S(S=s) := \frac{N_s}{N}.$$

Then we estimate the (population) joint by

$$\boxed{\widehat{\mathbb{P}}(Y=1, S=s) := \widehat{\mathbb{E}}_{X|S=s}\big[\widehat{\eta}(X,s)\big]\ \widehat{\mathbb{P}}_S(S=s)}$$

This equality holds by the law of total expectation and the definition of η:

$$\mathbb{P}(Y=1 \mid S=s) = \mathbb{E}[\,Y \mid S=s\,] = \mathbb{E}[\,\mathbb{E}[Y \mid X, S=s] \mid S=s\,] = \mathbb{E}[\,\eta(X,s) \mid S=s\,].$$

Empirical Unfairness (Definition)

For any classifier g, an estimator $\hat{\eta}$ based on the labeled dataset $\mathcal{D}_n$, and an unlabeled sample $\mathcal{D}_N$, the **empirical unfairness** is defined as

$$\hat{\Delta}(g, \mathbb{P}) := \left| \frac{\widehat{\mathbb{E}}_{X|S=1}\big[\hat{\eta}(X,1)\, g(X,1)\big]}{\widehat{\mathbb{E}}_{X|S=1}\big[\hat{\eta}(X,1)\big]} - \frac{\widehat{\mathbb{E}}_{X|S=0}\big[\hat{\eta}(X,0)\, g(X,0)\big]}{\widehat{\mathbb{E}}_{X|S=0}\big[\hat{\eta}(X,0)\big]} \right|.$$

Here, $\widehat{\mathbb{E}}_{X|S=s}[f(X)] = \frac{1}{N_s} \sum_{i=n+1}^{n+N} f(X_i)\, \mathbf{1}\{S_i = s\}$ *and* $\widehat{\mathbb{P}}_S(S = s) = \frac{N_s}{N}$ *are computed from the unlabeled dataset* $\mathcal{D}_N$*.*

Key Remark: The empirical unfairness $\hat{\Delta}(g, \mathbb{P})$ is *data-driven* and does not involve any unknown population quantities.

Estimation of θ

Recall that the EO–optimal classifier g^* can be written in terms of a parameter θ^*. Since θ^* and the true distribution $\mathbb{P}$ are unknown, we define the **empirical plug-in classifier** $\hat{g}_\theta$ by substituting empirical estimates:

$$\hat{g}_\theta(x,1) = \mathbf{1}\left\{1 \le \hat{\eta}(x,1)\left(2 - \frac{\theta}{\widehat{\mathbb{P}}(Y=1,S=1)}\right)\right\}, \quad \hat{g}_\theta(x,0) = \mathbf{1}\left\{1 \le \hat{\eta}(x,0)\left(2 + \frac{\theta}{\widehat{\mathbb{P}}(Y=1,S=0)}\right)\right\}.$$

We then estimate θ^* via

$$\hat{\theta} \in \arg\min_{\theta\in[-2,2]} \hat{\Delta}(\hat{g}_\theta, \mathbb{P}).$$

Why [-2,2]?
This interval ensures that the thresholds in $\hat{g}_\theta$ remain positive (i.e., well-defined), since $2 \pm \theta/\widehat{\mathbb{P}}(Y=1, S=s) > 0$ implies $|\theta| \le 2$.

Consistency

Theorem (Consistency of the plug-in rule). As $n, N \to \infty$ with $\hat{\eta}$ trained on $\mathcal{D}_n$ independently of $\mathcal{D}_N$,

$$\underbrace{\mathbb{E}[\Delta(\hat{g}, \mathbb{P})] \to 0}_{\text{asymptotically fair}} \quad \text{and} \quad \underbrace{\mathbb{E}[\mathcal{R}(\hat{g})] \to \mathcal{R}(g^*)}_{\text{asymptotically optimal}}.$$

- **Assumptions**

  1. **Regression consistency (A1)**
     $\mathbb{E}[\,|\hat{\eta}(X,S) - \eta(X,S)|\,] \to 0.$
  2. **No mass at thresholds / continuity (A2)**
     For $s \in \{0,1\}$, the law of $\eta(X,s)$ has no point mass at the EO thresholds; small neighborhoods have vanishing probability.
  3. **LLN on unlabeled sample (A3)**
     Empirical conditionals from $\mathcal{D}_N$ converge: $\widehat{\mathbb{E}}_{X|S=s}[f(X)] \to \mathbb{E}_{X|S=s}[f(X)]$ for bounded f.
  4. **Shift identification (A4)**
     $\Theta = [-2,2]$, and the population EO gap $\phi(\theta) := \mathrm{TPR}_1(g_\theta) - \mathrm{TPR}_0(g_\theta)$ has a unique root θ^*.

Consistency Proof Sketch (1/2)

Step 1. Population target. Let g_θ be the EO-threshold rule obtained by plugging the *true* quantities η and $\mathbb{P}(Y{=}1, S{=}s)$. Define the population EO gap

$$\phi(\theta) \;:=\; \Delta(g_\theta, \mathbb{P}) = \big|\mathrm{TPR}_1(g_\theta) - \mathrm{TPR}_0(g_\theta)\big|.$$

By shift identification, ϕ has a unique root θ^* and $\phi(\theta^*) = 0$.

Step 2. Empirical objective. Define the data-driven gap

$$\hat{\phi}(\theta) \;:=\; \hat{\Delta}(\hat{g}_\theta, \mathbb{P}),$$

where $\hat{g}_\theta$ uses $\hat{\eta}$ and $\widehat{\mathbb{P}}(Y{=}1, S{=}s)$ (from $\mathcal{D}_n$ and $\mathcal{D}_N$ respectively).

Consistency Proof Sketch (2/3)

Step 3. Uniform convergence of the EO gap.

$$\boxed{\sup_{\theta \in [-2,2]} \left| \hat{\phi}(\theta) - \phi(\theta) \right| \xrightarrow{p} 0}$$

Sketch. (i) Replace population conditionals by unlabeled empirical ones: LLN on $\mathcal{D}_N$ (**A3**).
(ii) Replace η by $\hat{\eta}$: regression consistency in L^1 (**A1**).
(iii) Handle the indicator discontinuity: no mass at the moving thresholds (**A2**) makes the boundary band negligible.
Compact parameter set $\Theta = [-2, 2]$ (**A4**) gives uniformity.

Step 4. Argmin consistency for the shift. By the M-estimation argmin theorem on compact Θ, uniform convergence (Step 3) and uniqueness (A4) imply

$$\hat{\theta} \in \arg\min_{\theta \in [-2,2]} \hat{\phi}(\theta) \quad \Longrightarrow \quad \hat{\theta} \xrightarrow{p} \theta^*.$$

Consistency Proof Sketch (3/3)

Step 5. Conclude fairness and risk consistency. Decision regions differ only where $\eta(X, S)$ lies in a vanishing band around the thresholds or where $\hat{\theta}$ deviates from θ^*:

$$\mathbb{P}\big(\hat{g}_{\hat{\theta}}(X, S) \neq g_{\theta^*}(X, S)\big) \;\to\; 0.$$

Hence

$$\Delta(\hat{g}, \mathbb{P}) = \Delta(\hat{g}_{\hat{\theta}}, \mathbb{P}) \;\to\; \Delta(g_{\theta^*}, \mathbb{P}) = 0,$$

and for the risk,

$$\big|\mathcal{R}(\hat{g}_{\hat{\theta}}) - \mathcal{R}(g_{\theta^*})\big| \;\leq\; \mathbb{E}[\,|2\eta(X, S) - 1|\,\mathbf{1}\{\hat{g}_{\hat{\theta}} \neq g_{\theta^*}\}] \;\to\; 0.$$

Takeaway. Uniform control of the EO gap $\Rightarrow$ $\hat{\theta} \to \theta^*$, and the plug-in classifier $\hat{g} = \hat{g}_{\hat{\theta}}$ becomes *asymptotically fair* and *risk-consistent*.

Conclusion

Conclusion

Contributions
We propose a label-efficient EO calibration method that leverages **unlabeled data** to learn a single group-dependent shift, avoiding retraining and heavy reliance on labels. Unlike prior approaches, we characterize the **EO-optimal rule in closed form** and establish strong **optimality and consistency guarantees**.

Experiment results
Across 5 datasets, our method consistently **reduces DEO with little or no loss in accuracy**. For example, with more unlabeled data (RF + Ours, fixed labeled budget $|\mathcal{D}_n| = 1/10$): COMPAS: ACC $0.68 \to 0.71$, DEO $0.07 \to 0.05$. Adult: ACC $0.79 \to 0.80$, DEO $0.06 \to 0.04$.