Fair Classification with Tail-Distribution Constraints

Joonhyuk Suh Stan Uryasev

Department of Applied Mathematics and Statistics, Stony Brook University

Introduction

Motivation In fairness classification, most approaches equalize the frequency of positive outcomes across groups, ensuring that similar proportions are classified above a decision threshold. However, in many high-stakes applications such as credit card issuance, credit limits, or recidivism risk assessment, outcomes depend not only on whether the threshold is exceeded but also on how far scores lie above it. Frequency-based parity (POE constraints) is valuable because it ensures equal opportunity of being classified positive, but in settings where extreme scores lead to harsher consequences, parity that also controls the tail impact becomes equally important. By incorporating CVaR and bPOE constraints, we can deal with the intensity of outcomes across groups, addressing disparities that frequency-based parity alone cannot resolve.

Idea We import three concepts—Probability of Exceedance (POE), Conditional Value-at-Risk (CVaR), and Buffered POE (bPOE)—into fairness-aware logistic regression to control both frequency (POE) and tail impact (CVaR and bPOE) at a fixed threshold.

Contributions

- Fairness criteria: *POE parity* and *bPOE parity* generalize statistical parity to thresholded risks.
- Optimization models: constrained and penalized logistic regression with POE, CVaR, and bPOE constraints.
- Evidence: COMPAS case study shows large fairness gains with modest accuracy deterioration.

Background

Definitions Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $t^+ = \max\{0, t\}$.

- Cumulative Distribution Function (CDF) :
$$F_X(x) = P(X \le x).$$
- Quantile function at α :
$$q_\alpha(X) = \min\{x \mid F_X(x) \ge \alpha\}.$$
- Probability of Exceedance (POE) at x:
$$p_x(X) = P(X > x),$$
- Conditional Value-at-Risk (CVaR) at α:
$$\bar{q}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 q_p(X)dp.$$
- Buffered Probability of Exceedance (bPOE) at x :
$$\bar{p}_x(X) = \min_{\lambda \ge 0} \mathbb{E}\left[\lambda(X - x) + 1\right]^+, \quad \text{for } x \ne \sup X.$$

Method

Setup $(\boldsymbol{X}, A, Y)$, where $\boldsymbol{X} \in \mathcal{X}$ is a feature vector, $A \in \mathcal{A}$ is a sensitive attribute, and $Y \in [0,1]$ is the label. The goal of fairness classification is to find a score function $f : \mathcal{X} \to \mathbb{R}$ that predicts Y given $\boldsymbol{X}$ while satisfying fairness conditions with a threshold $z \in \mathcal{Z}$. $\boldsymbol{w}_s$ is a shared coefficient vector and w_0^a is a subgroup-specific intercept for each sensitive group $a \in \mathcal{A}$.

Fairness parities

- POE parity : a score function f satisfies POE parity if
$$p_z(f(\boldsymbol{X}) \mid A = a) = p_z(f(\boldsymbol{X})) \quad \text{for all } a \in \mathcal{A} \text{ and } z \in \mathcal{Z}.$$
.
- bPOE parity : a score function f satisfies bPOE parity if
$$\bar{p}_z(f(\boldsymbol{X}) \mid A = a) = \bar{p}_z(f(\boldsymbol{X})) \quad \text{for all } a \in \mathcal{A} \text{ and } z \in \mathcal{Z}.$$

Model Logistic regression with group-specific intercepts:
$$P(\hat{Y} = 1 | \boldsymbol{X}, A = a) = \sigma\big(\boldsymbol{w}_s^\top \boldsymbol{X} + w_0^a\big), \qquad \sigma(t) = \tfrac{1}{1+e^{-t}}.$$

Optimization problems with two sensitive groups Let negative logistic log-likelihood function $L(\boldsymbol{w}, \boldsymbol{X})$ for a dataset $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^J$ is given by:
$$L(\boldsymbol{w}, \boldsymbol{X}) = \frac{1}{J}\sum_{i=1}^{J}\left[-Y_i \boldsymbol{w}^T \boldsymbol{X}_i + \log(1 + e^{\boldsymbol{w}^T \boldsymbol{X}_i})\right].$$
and $\boldsymbol{w} = \boldsymbol{w}_s^\top X + \sum_{a \in \mathcal{A}} w_0^a$

- *Problem 1 : POE-constrained*
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^n} \;& L(\boldsymbol{w}, \boldsymbol{X}) \\ \text{s.t.} \;& p_z(\boldsymbol{w}^T \boldsymbol{X} \mid A = 1) \le \mu \\ & p_{-z}(-\boldsymbol{w}^T \boldsymbol{X} \mid A = 2) \le 1 - \mu \end{aligned}$$
- *Problem 2 : Penalized objective*
$$\min_{\boldsymbol{w} \in \mathbb{R}^n} \; L(\boldsymbol{w}, \boldsymbol{X}) - \lambda(\mathbb{E}[\ell(\boldsymbol{w}^T \boldsymbol{X}) \mid A = 1] - \mathbb{E}[\ell(\boldsymbol{w}^T \boldsymbol{X}) \mid A = 2])$$
- *Problem 3 : CVaR-constrained*
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^n} \;& L(\boldsymbol{w}, \boldsymbol{X}) \\ \text{s.t.} \;& \bar{q}_\nu(\boldsymbol{w}^T \boldsymbol{X} \mid A = 1) \le b \\ & \bar{q}_{1-\nu}(-\boldsymbol{w}^T \boldsymbol{X} \mid A = 2) \le c \end{aligned}$$
- *Problem 4 : CVaR + mean-shift*
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^n} \;& L(\boldsymbol{w}, \boldsymbol{X}) + \gamma \mathbb{E}[\boldsymbol{w}^T \boldsymbol{X} \mid A = 2] \\ \text{s.t.} \;& \bar{q}_\nu(\boldsymbol{w}^T \boldsymbol{X} \mid A = 1) \le b \\ & \bar{q}_\nu(\boldsymbol{w}^T \boldsymbol{X} \mid A = 2) \le b \end{aligned}$$
- *Problem 5 : bPOE-constrained*
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^n} \;& L(\boldsymbol{w}, \boldsymbol{X}) \\ \text{s.t.} \;& \bar{p}_b(\boldsymbol{w}^T \boldsymbol{X} \mid A = 1) \le \nu \\ & \bar{p}_c(-\boldsymbol{w}^T \boldsymbol{X} \mid A = 2) \le 1 - \nu \end{aligned}$$

Data

Source ProPublica COMPAS (Broward County, FL): demographics (race, age, gender), criminal history, COMPAS risk scores; standard benchmark for fairness studies.

Setup Observations $(\boldsymbol{X}, A, Y)$ with two sensitive groups: White (A=1) and Black (A=2). Features: age, decile score, priors count. Label: criminal recidivism.

Preprocessing Features standardized (z-score). Design includes group intercept indicators.

Calibration (unconstrained model) Proportion of Y=1: $\mu = 0.4704$; VaR at level μ: $z = 0.2593$; complement $\nu = 1 - \mu = 0.5296$; CVaR at level μ: $b = 0.9301$; CVaR at level $1-\nu$: $c = 0.0327$. Sensitivity-tuned hyperparameters: $\lambda = 0.0296$, $\gamma = 0.0431$.

Training framework All formulations solved with the PSG (Portfolio Safeguard) optimization framework.

Results (Formulations)

Problem 1: Logistic regression with POE constraints
Parameters: z=0.2593, μ=0.4704.
POE at z: Overall: 0.4719, A=1: 0.4703, A=2: 0.4728 (gap 0.0025).

Problem 2: Logistic regression with penalized objective
Parameters: z=0.2593, λ=0.0296.
POE at z: Overall: 0.4740, A=1: 0.4743, A=2: 0.4736 (gap 0.0007).

Problem 3: Logistic regression with CVaR constraints
Parameters: ν=0.5296, b=0.9301, c=0.0327.
bPOE at b: Overall: 0.4697, A=1: 0.4704, A=2: 0.4691 (gap 0.0014).

Problem 4: Logistic regression with CVaR + mean-shift
Parameters: ν=0.5296, b=0.9301, γ=0.0431.
bPOE at b: Overall: 0.4707, A=1: 0.4704, A=2: 0.4704 (gap 0.0000).

Problem 5: Logistic regression with bPOE constraints
Parameters: ν=0.4704, b=0.9301, c=0.0327.
bPOE at b: Overall: 0.4702, A=1: 0.4715, A=2: 0.4693 (gap 0.0022).

Takeaway All five formulations reduce group disparity at the target threshold; Problem 4 enforces exact parity, while Problem 1, 2, 3, and 5 achieve near-parity with small differences.

Results (Metrics & Comparison)

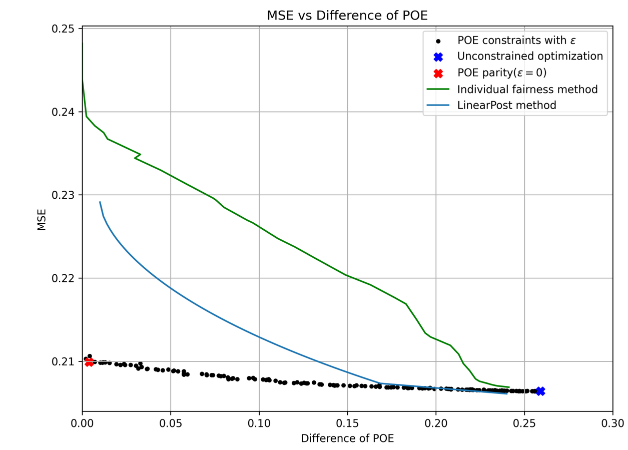

Fig 1 Fairness–accuracy frontiers with POE of Problem 1: Mean-squared error (MSE; y-axis) vs. absolute POE gap between groups (x-axis). Black dots: POE-constrained models across tolerance ϵ; red "×": exact POE parity (ϵ=0); blue "×": unconstrained model; green curve: individual-fairness model; Blue curve: LinearPost model.

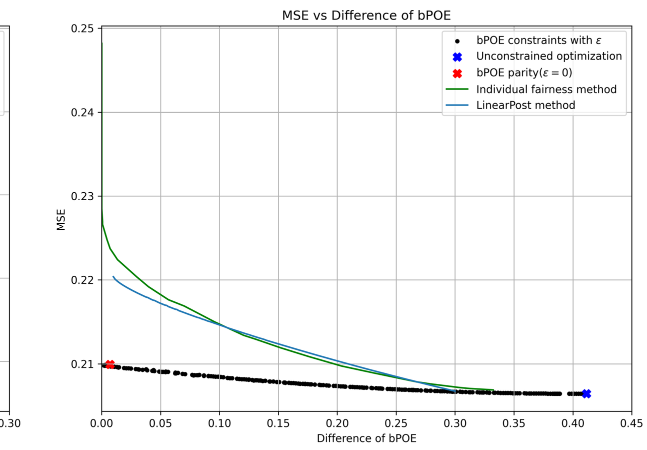

Fig 2 Fairness–accuracy frontiers with bPOE of Problem 5: MSE (y-axis) vs. absolute bPOE gap between groups (x-axis). Black dots: bPOE-constrained models across ϵ; red "×": exact bPOE parity (ϵ=0); blue "×": unconstrained model; green curve: individual-fairness model; Blue curve: LinearPost model.

Problem	MSE	Accuracy	TPR	Precision	AUC	bAUC
Original	0.206429	0.682266	0.663713	0.661847	0.739454	0.392172
Problem 1	0.209958	0.667109	0.646395	0.646135	0.729256	0.369757
Problem 2	0.210469	0.663698	0.643174	0.642397	0.727096	0.368840
Problem 3	0.219704	0.662941	0.641965	0.641707	0.725911	0.366332
Problem 4	0.211001	0.666161	0.645187	0.645187	0.729175	0.372549
Problem 5	0.219697	0.662941	0.641965	0.641707	0.725925	0.366429

Fig 3 Key metrics for Original and from Problem 1 to Problem 5 (MSE, Accuracy, TPR, Precision, AUC, bAUC), showing fairness–accuracy trade-offs across formulations.

Takeaway

- **Fig 1** and **Fig 2** compare POE-constrained and bPOE-constrained model (black dots) with the individual fairness model (green curve) and LinearPost model (blue curve). POE and bPOE-based approaches achieve lower fairness gaps at lower increases in MSE rather than other models.
- In **Fig 3**, while all proposed formulations achieve fairness constraints to some extent, Problems 1 and 4 demonstrate superior performance in balancing fairness with classification accuracy, as reflected in their AUC and bAUC improvements.

Conclusion

We propose a framework for fair classification that incorporates both POE and bPOE. POE parity ensures fairness in terms of frequency, i.e., how often individuals in each group exceed a decision threshold. bPOE parity extends this by controlling tail impact, i.e., how severe the exceedances are when they occur. Empirically on COMPAS, POE- and bPOE-based formulations reduce group disparities with competitive accuracy.

11th Annual Bloomberg-Columbia Machine Learning in Finance Conference 2025