Optimal Fair Aggregation of Crowdsourced Noisy Labels using Demographic Parity Constraints

Gabriel Singer*,1 Samuel Gruffaz*,1 Olivier Vo Van[2] Nicolas Vayatis[1] Argyris Kalogeratos[1]

*Equal contribution [1]Univ. Paris-Saclay, ENS Paris-Saclay, Centre Borelli [2]SNCF

ICML International Conference On Machine Learning

Problem setup and fairness criterion

Observe R noisy crowd labels

$$\bar{Y}_{1:R} = (Y_1, \dots, \bar{Y}_R) \in \{0,1\}^R$$

for a latent binary label $Y \in \{0,1\}$, features $X \in \mathcal{X}$, and sensitive attribute $A \in \{0,1\}$.

The **demographic-parity gap** of an aggregation rule ϕ is

$$\Delta_{\mathrm{DP}}(\bar{Y}^\phi) := \left|\mathbb{P}(\bar{Y}^\phi = 1 \mid A = 1) - \mathbb{P}(\bar{Y}^\phi = 1 \mid A = 0)\right|.$$

An aggregation rule ϕ is ε-fair when

$$\Delta_{\mathrm{DP}}(\bar{Y}^\phi) \le \varepsilon.$$

Two reference aggregation rules

Majority vote

$$\mathbb{1}\left\{\sum_{r=1}^{R} \bar{Y}_r \ge \frac{R}{2}\right\}.$$

Uses only the hard crowd labels.

Bayesian vote

$$\mathbb{1}\left\{\mathbb{P}(Y = 1 \mid \bar{Y}_{1:R}, X, A) \ge \frac{1}{2}\right\}.$$

Uses the full posterior information.

Takeaway. We compare how standard crowd aggregation rules distort or preserve demographic parity.

Non-asymptotic fairness inequality

For $\phi \in \{\phi^\star, \phi^{\mathrm{MV}}\}$ and any R:

$$\left|\Delta_{\mathrm{DP}}(\bar{Y}_R^\phi) - \Delta_{\mathrm{DP}}(Y)\right| \le \sum_{a \in \{0,1\}} \mathbb{E}\left[e^{-RK_\phi(a,X)} \mid A = a\right].$$

The right-hand side decays exponentially in R: the fairness gap of $\bar{Y}_R^\phi$ converges to the bias already present in the latent truth Y.

ε-Demographic Parity Gap of the Majority vote

$$\Delta_{\mathrm{DP}}(\bar{Y}_R^{\phi^{\mathrm{MV}}}) \le \varepsilon_R \sum_{r=1}^{R} \Delta_{\mathrm{DP}}(\bar{Y}_r),$$

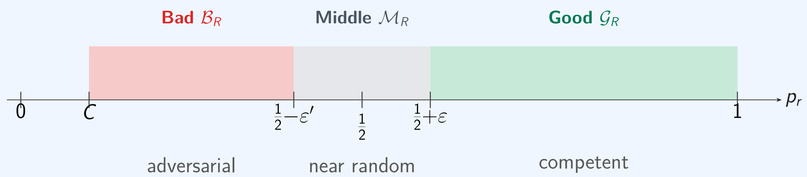
$$\varepsilon_R := \frac{\eta}{\min\left\{\sqrt{V_R(0)}, \sqrt{V_R(1)}\right\}},$$

with $\eta \simeq 0.4688$.

Interpretable crowdsourcing conditions

Rank annotators by accuracy

$$p_r := \mathbb{P}(\bar{Y}_r = Y \mid X, A) \in [0,1].$$



Majority vote:

$$\liminf_{R\to\infty} \left\lceil \frac{|\mathcal{G}_R|}{R}(1+2\varepsilon) + 2C\,\frac{|\mathcal{B}_R|}{R} \right\rceil > 1 \qquad \text{a.s.}$$

Bayesian vote:

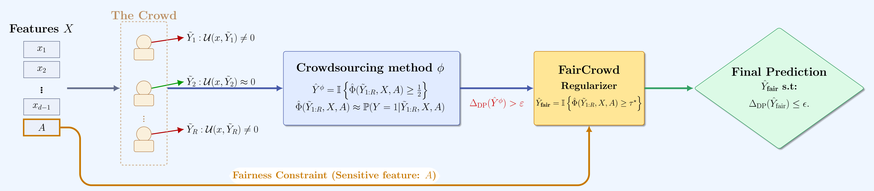
$$\sum_{r=1}^{\infty} \left|p_r - \frac{1}{2}\right|^2 = \infty \qquad \text{a.s.}$$

Under these conditions:

$$\lim_{R\to\infty} \Delta_{\mathrm{DP}}(\bar{Y}_R^\phi) = \Delta_{\mathrm{DP}}(Y), \qquad \phi \in \{\phi^\star, \phi^{\mathrm{MV}}\}.$$

Takeaway. Large crowds recover, not erase, ground-truth bias.

FairCrowd: optimal ε-fair post-processing



Closed-form formula: boundary shift based on fairness trade-off

Instead of modifying the crowd model, FairCrowd solves

$$\min_{\phi \in \mathcal{G}_\varepsilon} \mathbb{P}(\bar{Y}^\phi \neq Y) \quad \text{subject to} \quad \Delta_{\mathrm{DP}}(\bar{Y}^\phi) \le \varepsilon.$$

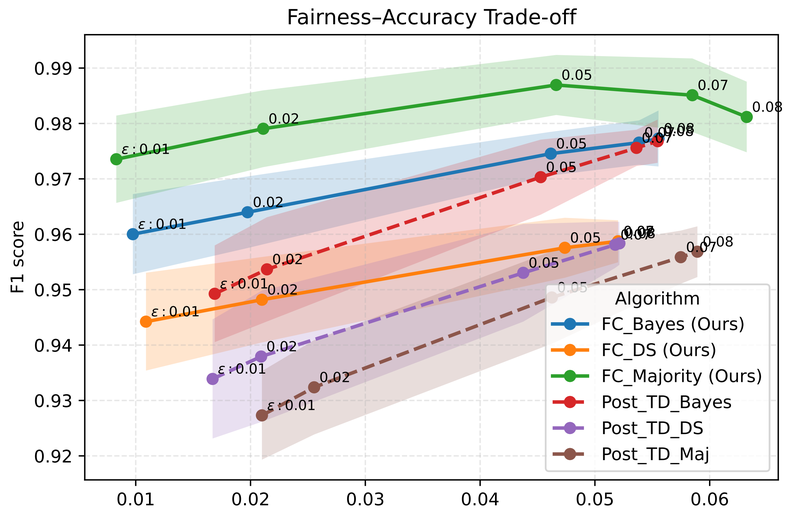
The closed-form optimal rule is a **group-dependent threshold shift**:

$$\phi^\star_{\beta^\star}(w, a) = \mathbb{1}\left\{\Phi_1(w, a) > \frac{\pi_a + (2a-1)\beta^\star}{2\pi_a}\right\},$$

where $\pi_a = \mathbb{P}(A = a)$ and $\beta^\star$ is the unique dual variable that saturates

$$\Delta_{\mathrm{DP}}(\bar{Y}^{\phi^\star}) = \varepsilon.$$

Experiments: fairness–accuracy trade-off



Jigsaw Toxicity. Solid: FairCrowd. Dashed: Post TD baselines. Curves sweep ε.

Empirical results

Accuracy: FairCrowd enforces the prescribed fairness constraint while preserving, and in some cases improving, predictive performance compared to FairTD and Post TD baselines.

Scalability. On large datasets, the post-processing step remains computationally negligible, with a running time below 3 seconds.

Main references

C. Denis, R. Elie, M. Hebiri, and F. Hu. *Fairness guarantees in multi-class classification with demographic parity.* JMLR, 25(130):1–46, 2024.

C. Gao, Y. Lu, and D. Zhou. *Exact exponent in optimal rates for crowdsourcing.* ICML, pp. 603–611, 2016.

S. Lazier, S. Thirumuruganathan, and H. Anahideh. *Fairness and bias in truth discovery algorithms.* arXiv:2304.12573, 2023.