Optimal Fair Aggregation of Crowdsourced Noisy Labels using Demographic Parity Constraints

Singer Gabriel, Gruffaz Samuel , Vo Van Olivier, Vayatis Nicolas, Kalogeratos Argyris

ICML 2026

Motivation

Two gaps in the literature, two research questions

G1. *Empirically*, naive aggregation amplifies aligned biases instead of cancelling them [Lazier '23].

→

Q1. Can we prove it *theoretically*, once and for all, and quantify when the consensus actually becomes fair?

G2. Existing fair classifiers [Denis '24, Xian '23, Zeng '22] assume *continuous* posteriors: they exclude the discrete crowdsourcing case. SOTA [Li '20]: heuristic, no theory, $>$ 30h on Jigsaw.

→

Q2. Can we build a *theoritically grounded* and *fast* post-processor that turns any aggregator into a strict ε-DP rule on *discrete* inputs?

Problem setup & Demographic Parity

Variables. $Y \in \{0,1\}$ latent; $\widetilde{Y}_{1:R} \in \{0,1\}^R$ noisy labels; features $X \in \mathcal{X}$; sensitive $A \in \{0,1\}$.

Aggregation. $\phi : \{0,1\}^R \times \mathcal{X} \times \{0,1\} \to \{0,1\}, \quad \widehat{Y}^{\phi} := \phi(\widetilde{Y}_{1:R}, X, A)$.

One-coin skill. $p_r(a,x) := \mathbb{P}(\widetilde{Y}_r = Y \mid X = x,\ A = a) \in [0,1]$.

Global DP gap

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

Population-level. $\widetilde{Y}$ *is* ε*-fair iff* $\Delta_{\mathrm{DP}} \le \varepsilon$.

Local DP gap

$Q_a(x) := \mathbb{P}(\widetilde{Y}{=}1 \mid X{=}x,\ A{=}a)$

$U(\widetilde{Y}, x) := \left|Q_1(x) - Q_0(x)\right|$

Pointwise disparity at a fixed x.

Theoretical analysis

Two natural aggregators

Majority Vote ϕ^{MV}

$$\phi^{\mathrm{MV}}(\widetilde{Y}_{1:R}, X, A) = \mathbb{1}\left\{\sum_{r=1}^{R} \widetilde{Y}_r \geq \frac{R}{2}\right\}$$

- no model of the crowd
- scalable, ubiquitous

Bayes Optimal $\phi^\star$

$$\phi^\star = \mathbb{1}\{\mathbb{P}(Y = 1 \mid \widetilde{Y}_{1:R}, X, A) \geq \tfrac{1}{2}\}$$

- minimises 0–1 risk
- needs annotator skills (or EM)

Starting point [Gao et al. 2016]:

$$\mathbb{P}\big(\widehat{Y}_R^{\phi} \neq Y \mid X = x, A = a\big) \;\leq\; \exp\big(-R\, K_\phi(a, x)\big).$$

Takeaway. Error decays exponentially in R.
We turn this into a *fairness-gap* bound.

From error decay to fairness convergence

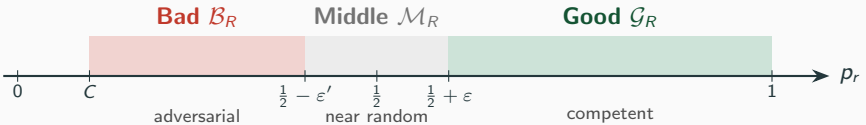
Proposition 1 (Non-asymptotic fairness bound)
For any $R \geq 1$ and $\phi \in \{\phi^\star, \phi^{\mathrm{MV}}\}$:

$$\left|\Delta_{\mathrm{DP}}(\widehat{Y}_R^{\phi}) - \Delta_{\mathrm{DP}}(Y)\right| \;\leq\; \sum_{a\in\{0,1\}} \mathbb{E}_{X|A=a}\left[e^{-R\,K_\phi(a,X)}\right].$$

Why it matters. Sufficient conditions of Gao et al. are opaque. We propose *interpretable* ones.

An interpretable partition of the crowd

Rank annotators by their accuracy $p_r := \mathbb{P}(\widetilde{Y}_r = Y \mid X, A) \in [0,1]$, then split:



When does $\Delta_{\mathrm{DP}}(\widehat{Y}_R^{\phi}) \xrightarrow{R\to\infty} \Delta_{\mathrm{DP}}(Y)$**?**

Majority Vote

$$\liminf_{R\to\infty} \left[\frac{|\mathcal{G}_R|}{R}(1+2\varepsilon) + 2C\,\frac{|\mathcal{B}_R|}{R} \right] > 1$$

A **majority of experts** *must outweigh adversaries.*

Bayesian Vote

$$\sum_{r=1}^{\infty} \left(p_r - \tfrac{1}{2}\right)^2 = \infty \quad \text{a.s.}$$

Just need **some annotators not to be pure noise**.

Takeaway. In the large-crowd regime, aggregation *inherits* the fairness of Y. The crowd cannot "erase" the biases at the limit.

What about the small-crowd regime?

A single annotator is biased by $\Delta_{\mathrm{DP}}(\widetilde{Y}_r)$. What does Majority Vote produce with few of them?

Proposition 2 (Small-crowd bound for Majority Vote)
With $V_R(a) = \sum_{i=1}^{R} l_i(a)(1 - l_i(a))$, *there exists* $\eta \approx 0.4688$ *such that*

$$\Delta_{\mathrm{DP}}\big(\widehat{Y}_R^{\phi^{\mathrm{MV}}}\big) \;\le\; \underbrace{\eta\Big(\min\{\sqrt{V_R(0)}, \sqrt{V_R(1)}\}\Big)^{-1}}_{\varepsilon(R)} \sum_{r=1}^{R} \Delta_{\mathrm{DP}}(\widetilde{Y}_r)$$

If biases align with $\Delta_{\mathrm{DP}}(\widetilde{Y}_r) \approx \delta > 0$, the aggregate bias grows as $\mathcal{O}(\sqrt{R})\,\delta$ until saturating at 1.

Why it matters. Majority Vote is a *bias accumulator*, not a filter, when R is small. $\Rightarrow$ post-processing needed.

The FairCrowd algorithm

The big picture

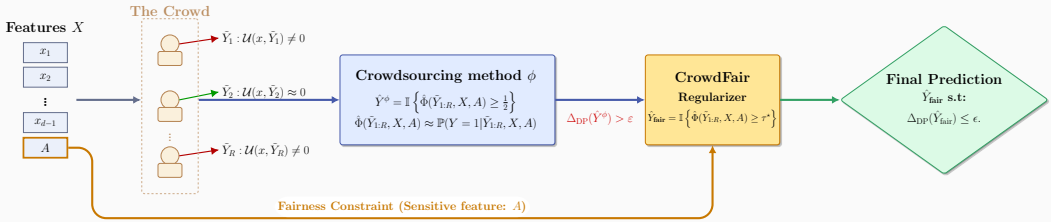


Figure 2: Our algorithm :)

Takeaway. FairCrowd is a *Post processing* step: it takes any aggregator's posterior $\widehat{\Phi}$ and shifts its decision threshold to enforce strict ε-DP.

Optimal ε-fair classifier on discrete inputs

Why a new theorem? [Denis et al. 2024] requires $t \mapsto \mathbb{P}\big(P_1^*(W,A) \le t \mid A{=}a\big)$ to be *continuous*. In crowdsourcing $W = \widetilde{Y}_{1:R} \in \{0,1\}^R$ is discrete — the c.d.f. has jumps, so we redo the analysis with sub-gradients and randomised classifiers.

Theorem 4.1 (binary case)
Let $s_a = 2a - 1$, $\pi_a = \mathbb{P}(A = a)$. The solution of $\min_{\phi \in \mathcal{G}_\varepsilon} \mathbb{P}(\widehat{Y}^\phi \neq Y)$ is

$$\phi^*_{\beta^*}(w,a) = \begin{cases} 1, & P_1^*(w,a) > \dfrac{\pi_a + s_a\beta^*}{2\pi_a}, \\ 0, & P_1^*(w,a) < \dfrac{\pi_a + s_a\beta^*}{2\pi_a}, \\ \omega_a, & \text{equality (randomisation at ties)}, \end{cases}$$

with $\beta^* = \arg\min_\beta \{L(\beta) + \varepsilon|\beta|\}$ and

$$L(\beta) = \sum_{a=0}^{1} \mathbb{E}_{W|A=a}\Big[\max_{k\in\{0,1\}} \big(\pi_a P_k^*(W,a) - \tfrac{\beta}{2} s_a s_k\big)\Big].$$

Experiments

Empirical validation of the theory

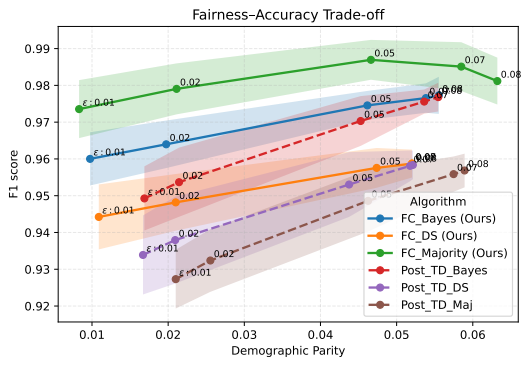


Figure 3: Performance comparison on the Jigsaw dataset. Solid lines correspond to our method (FC), while dashed lines indicate competing approaches. Shaded regions denote the variance across 10 independent runs using different test set (60%).

Results on Jigsaw Toxicity

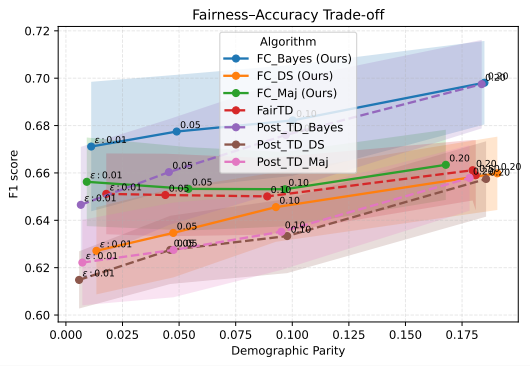


Figure 4: Crowd Judgement dataset. Solid lines correspond to our method (FC), while dashed lines indicate competing approaches. Shaded regions denote the variance across 10 independent runs using different test set (60%).