Measuring Model Fairness under Noisy Covariates: A Theoretical Perspective

Flavien Prost, Pranjal Awasthi, Nick Blumm, Aditee Kumthekar, Trevor Potter, Li Wei, Xuezhi Wang, Ed H. Chi, Jilin Chen, Alex Beutel

Google

Problem

Measuring a group fairness metric such as statistical parity or equal opportunity under noisy data.

Motivation

- Toxicity classifier (*y*) for comments.
- Goal: Measure fairness across demographics (*l=0,1*) but only for specific topic (*v=1*).
- → **Ideally,** we want to evaluate *Conditional Statistical Parity** as:

$$G_{SP} = \mathbb{P}[y=1|v=1,\ell=0] - \mathbb{P}[y=1|v=1,\ell=1]$$

- **Challenge:** (*y, l, v*) not jointly observable.

Problem Formulation

- (*y, l, v*) not jointly observable. What about using a topic classifier ($\hat{v}$)?

$$\hat{G}_{SP} = \mathbb{P}[y=1|\hat{v}=1,\ell=0] - \mathbb{P}[y=1|\hat{v}=1,\ell=1]$$

- **Objective**: Provide upper bounds for $|G_{SP} - \hat{G}_{SP}|$

Contributions

- We characterize a variety of conditions under which the estimation error above can be bounded.
- These bounds depend on the precision/recall of the classifier $\hat{v}$ as well as the joint correlations between y,v, and $\hat{v}$.

* Results apply to equal opportunity too.

Bounds Based only on Classifier Performance

$$\mathbb{P}[v=1|\hat{v}=1,\ell=l] = 1 - p_l$$
$$\mathbb{P}[\hat{v}=1|v=1,\ell=l] = 1 - r_l$$

Bound A

Theorem 5.1: If for any $\ell \in \{0,1\}$ the precision and recall of the proxy $\hat{v}$ is at least $1-\gamma_A$,

$$|G - \hat{G}| \leq 2 \cdot \gamma_A.$$

Question: Can we derive a better bound by assuming *some structure*?

Alternate Bounds based on General Correlation

- **Idea: Assumptions on parameters (y, v, $\hat{v}$, l).**

$\mathbb{P}[y=1|v=0,\hat{v}=1,\ell=l]$	$\mathbb{P}[y=1|v=1,\hat{v}=1,\ell=l]$
$\mathbb{P}[y=1|v=0,\hat{v}=0,\ell=l]$	$\mathbb{P}[y=1|v=1,\hat{v}=0,\ell=l]$

Table 1: Value of the outcome y over the confusion matrix of $v, \hat{v}$, conditioned on group $\ell = l$.

Case B1

Condition "Closeness of Diagonals": There exists ϵ_{B1} such that

$$\left|\Pr[y=1|v=1,\hat{v}=0,\ell=0] - \Pr[y=1|v=0,\hat{v}=1,\ell=0]\right| \leq \epsilon_{B1}$$
$$\left|\Pr[y=1|v=1,\hat{v}=0,\ell=1] - \Pr[y=1|v=0,\hat{v}=1,\ell=1]\right| \leq \epsilon_{B1}$$

Bound B1

Theorem: Let ϵ_{B1} be such that that closeness of diagonal condition holds with ϵ_{B1} and that $|r_0 - p_0|, |r_1 - p_1|$ are bounded by γ_{B1}, then

$$|G - \hat{G}| \leq 2(\gamma_{B1} + \epsilon_{B1}).$$

Case B2

Condition "Model Closeness": There exists g, ϵ_{B2} such that for all b, c:

$$\Pr[y=1|v=b,\hat{v}=c,\ell=1] = \Pr[y=1|v=b,\hat{v}=c,\ell=0] + g \pm \epsilon_{B2}.$$

Bound B2

Theorem: Let $\gamma_{B2}, \epsilon_{B2}, g$ be such that the model closeness holds with ϵ_{B2} and that $|r_0 - r_1|, |p_0 - p_1|$ are bounded by γ_{B2}. Then we have that

$$|G - \hat{G}| \leq 2 \cdot \gamma_{B2} + 3 \cdot \epsilon_{B2}.$$

Refined Bound

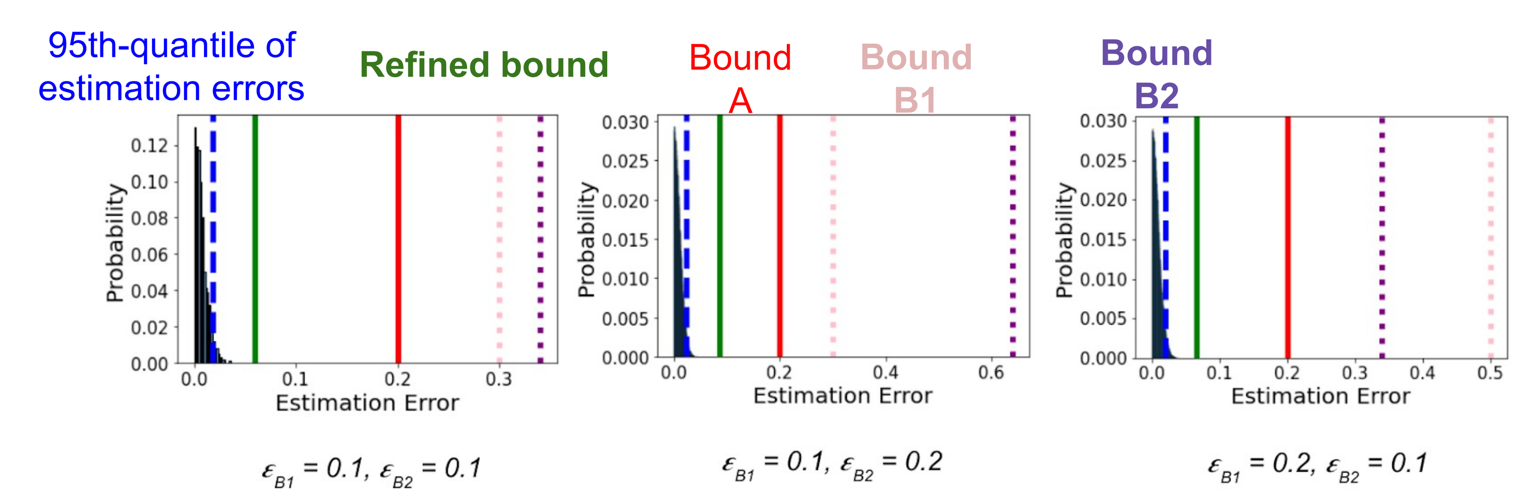
- In isolation, previous bounds might be loose → **Let's combine them!**

Theorem: Let γ_A, $(\gamma_{B1}, \epsilon_{B1})$ and $(\gamma_{B2}, \epsilon_{B2})$ be the errors up to which conditions A, B1 and B2 above hold. Then

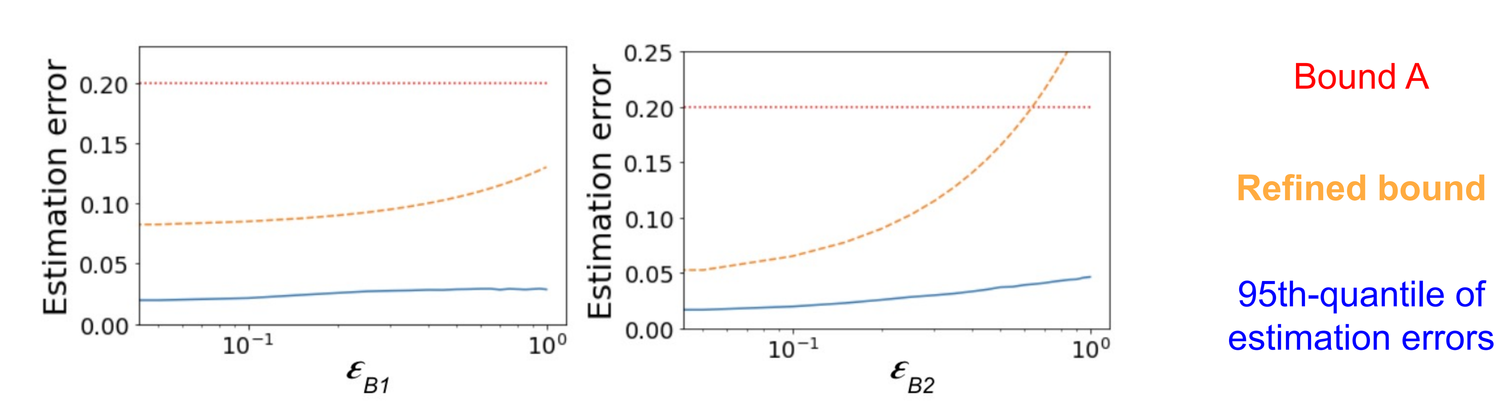
$$|G - \hat{G}| \leq 2\min(\gamma_A, \gamma_{B1}, \gamma_{B2}) + \epsilon_{B2} \cdot (2\gamma_A + \gamma_{B1}) + \epsilon_{B1} \cdot \gamma_{B1}$$

- Note: Linear in **γ** and quadratic in ε .

Simulations



- **Lesson: Refined bound dominates other ones.**



- **Lesson: Even weak assumptions on ε are enough to significantly reduce the bound.**