Internship Proposal (Research): "A mathematical analysis of positive discrimination mechanisms in multi-dimensions"

Keywords: Fairness, Implicit bias, Selection problems, Positive discrimination

Host lab & team: Laboratoire d'Informatique de Grenoble (LIG), Grenoble, France; POLARIS team

Supervisors:
Patrick Loiseau (Inria & Ecole Polytechnique) – http://lig-membres.imag.fr/loiseapa/ – patrick.loiseau@inria.fr
Bary Pradelski (CNRS) – https://barypradelski.com/ – bary.pradelski@cnrs.fr

Background

Many selection problems such as hiring or college admission are subject to discrimination [1], where the outcomes for certain individuals are negatively correlated with their membership in salient demographic groups defined by attributes like gender or race. Over the past two decades, implicit bias—that is an unconscious negative perception of the members of certain demographic groups—has been put forward as a key factor in explaining this discrimination [5]; and positive discrimination mechanisms (such as the Rooney rule that imposes that at least one candidate from the underrepresented group must be interviewed for a given position) were introduced to mitigate the effect of discrimination on underrepresented groups.

Yet, the mathematical analysis of positive discrimination mechanisms under implicit bias started only recently. [6] introduced a model of implicit bias and showed that, contrary to conventional wisdom, the Rooney Rule not only reduces discrimination but also increases the joint utility of short-listed candidates in many selection problems. [3] then introduced a *generalized Rooney Rule* for ranking problems imposing that a fixed proportion of candidates from the underprivileged group are represented in the top-k for all k. They showed that for bounded utility distributions, the generalized Rooney Rule can recover almost all the utility lost due to implicit bias.

Goal of the internship

Prior works consider only the case where the population is split into two groups (G_a, G_b), defined by a single sensitive attribute; e.g., this can refer to gender or race but not both dimensions simultaneously. In this internship, we propose to study a more general (and realistic) setting where a candidate has multiple sensitive attributes.

To fix ideas, consider the example of gender—men (a) or women (b)—and race—white (A) or black (B)—; and assume w.l.o.g. that b and B are the underprivileged groups. Then we consider the following model: Each candidate $i \in \{1, \cdots, m\}$ has a true latent utility $w_i \in \mathbb{R}^+$, which is the utility they would generate if hired. On the other hand, the observed utility $\hat{w}_i \leq w_i$ is the selector's estimate of w_i. The selector selects $n \leq m$ candidates with the highest observed utility (as he aims to maximize the sum of utilities). Implicit bias is modeled via multiplicative factors, that is, $\beta_{aB}, \beta_{bA}, \beta_{bB}$ (we set $\beta_{aA} = 1$ for completeness) where the observed utility for a candidate i of type $\theta \in \{a, b\}, \sigma \in \{A, B\}$ is given by

$$\hat{w}_i = \beta_{\theta\sigma} \cdot w_i. \tag{1}$$

We will consider different natural ways to model the relations between different implicit biases and different natural candidates for the extension of the Rooney Rule to this setting. The intern will then investigate the proposed model with the goal of characterizing mathematically the effect of the Rooney Rule extension; in particular starting with the following questions: Can we find a simple rule that achieves almost all the utility lost due to implicit bias (with some distributional assumptions)? How useful is it to know the bias parameters β? Can we generalize this to more than two sensitive attributes and how can we handle the curse of dimensionality? (Note that, in a related context, [2] analyze the two-dimensional categorization and its implications on the existence on self-financing taxes. In their context, 'biases' are additive and a simple independent rule does not suffice to overcome them.) In a second step, we will introduce noise in $\hat{w}_i$ as in [4], study risk-averse decision makers, and perform numerical simulations on public datasets.

Expected ability of the student and additional information

A strong background in probability and statistics is necessary; along with interest in modeling and in fairness issues. The internship is part of the Explainable and Responsible AI chair of the MIAI@Grenoble Alpes institute and will be hosted in the POLARIS team, a joint team between Inria and LIG (Grenoble CS lab). It may be continued as a PhD. For more information, please contact patrick.loiseau@inria.fr and bary.pradelski@cnrs.fr.

References

[1] M. Bertrand and S. Mullainathan. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination". In: *American Economic Review* 94.4 (Sept. 2004), pp. 991–1013.

[2] J.-P. Carvalho and B. S. R. Pradelski. *Identity and underrepresentation.* working paper. 2018.

[3] L. E. Celis, A. Mehrotra, and N. K. Vishnoi. "Interventions for ranking in the presence of implicit bias". In: *Conference on Fairness, Accountability, and Transperence (FAT* '20)* 1 (2020).

[4] V. Emelianov, N. Gast, K. P. Gummadi, and P. Loiseau. "On Fair Selection in the Presence of Implicit Variance". In: *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20)*. 2020, 649–675.

[5] A. Greenwald and L. Krieger. "Implicit Bias: Scientific Foundations". In: *California Law Review* 94 (July 2006), p. 945.

[6] J. M. Kleinberg and M. Raghavan. "Selection problems in the presence of implicit bias". In: *ITCS* 94.33 (2018), pp. 1–17.