Metric-Free Individual Fairness in Online Learning*

Yahav Bechavod[1], Christopher Jung[2], and Zhiwei Steven Wu[3]

[1]Hebrew University
[2]University of Pennsylvania
[3]Carnegie Mellon University

We study an online learning problem subject to the constraint of individual fairness, which requires that similar individuals are treated similarly. Unlike prior work on individual fairness, we do not assume the similarity measure among individuals is known, nor do we assume that such measure takes a certain parametric form. Instead, we leverage the existence of an *auditor* who detects fairness violations without enunciating the quantitative measure. In each round, the auditor examines the learner's decisions and attempts to identify a pair of individuals that are treated unfairly by the learner. We provide a general reduction framework that reduces online classification in our model to standard online classification, which allows us to leverage existing online learning algorithms to achieve sub-linear regret and number of fairness violations. Surprisingly, in the stochastic setting where the data are drawn independently from a distribution, we are also able to establish PAC-style fairness and accuracy generalization guarantees ([3]), despite only having access to a very restricted form of fairness feedback. Our fairness generalization bound qualitatively matches the uniform convergence bound of [3], while also providing a meaningful accuracy generalization guarantee. Our results resolve an open question by [2] by showing that online learning under an unknown individual fairness constraint is possible even without assuming a strong parametric form of the underlying similarity measure.

References

[1] Y. Bechavod, C. Jung, and Z. S. Wu. Metric-free individual fairness in online learning. *arXiv preprint arXiv:2002.05474*, 2020.

[2] S. Gillen, C. Jung, M. J. Kearns, and A. Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2605–2614, 2018.

[3] G. N. Rothblum and G. Yona. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674, 2018.

[*] The full version of this paper is available on arXiv [1]