Causal Feature Discovery through Strategic Modification

Yahav Bechavod* Katrina Ligett† Zhiwei Steven Wu‡ Juba Ziani§

June 13, 2020

Abstract

We consider an online regression setting in which individuals adapt to the regression model: arriving individuals may access the model throughout the process, and invest strategically in modifying their own features so as to improve their assigned score. We find that this strategic manipulation may help a learner recover the causal variables, in settings where an agent can invest in improving impactful features that also improve his true label. We show that even simple behavior on the learner's part (i.e., periodically updating her model based on the observed data so far, via least-square regression) allows her to simultaneously i) accurately recover which features have an impact on an agent's true label, provided they have been invested in significantly, and ii) incentivize agents to invest in these impactful features, rather than in features that have no effect on their true label.

This submission is a research paper. The full paper can be found at https://arxiv.org/abs/2002.07024.

*School of Computer Science and Engineering, The Hebrew University. Email: yahav.bechavod@cs.huji.ac.il.

†School of Computer Science and Engineering, The Hebrew University. Email: katrina@cs.huji.ac.il.

‡Computer Science and Engineering Department, University of Minnesota. Email: zsw@umn.edu.

§Warren Center for Network and Data Sciences, University of Pennsylvania. Email: jziani@seas.upenn.edu.