Multigroup Robustness (Abstract)

Lunjia Hu[*] Charlotte Peale[†] Judy Hanwen Shen[‡]

To address the shortcomings of real-world datasets, robust learning algorithms have been designed to overcome arbitrary and indiscriminate data corruption. However, practical processes of gathering data may lead to patterns of data corruption that are localized to specific partitions of the training dataset. Motivated by critical applications where the learned model is deployed to make predictions about people from a rich collection of overlapping subpopulations, we initiate the study of *multigroup robust* algorithms whose robustness guarantees for each subpopulation only degrade with the amount of data corruption *inside* that subpopulation. We propose a new definition of robustness for binary-labeled prediction tasks, which, at a high-level, guarantees that for a set of subpopulations, every group's accuracy-in-expectation is robust to dataset corruptions to points outside of the group (Figure 1).

We observe that classic agnostic learning algorithms such as empirical risk minimization over the benchmark class can fail to satisfy multigroup robustness even for extremely simple families of sub-populations, and empirically demonstrate that several standard models for classification fail to preserve multigroup robustness under simple label-flipping and data addition attacks on the Adult Income Dataset.

In light of these successful attacks, new ideas are necessary to develop algorithms that can match the performance guarantees of standard learning approaches while being multigroup robust. Towards this goal, we present a set of sufficient conditions for an algorithm to satisfy multigroup robustness. We reach these sufficient properties by connecting the notion of multigroup robustness to the existing notion of *multiaccuracy*, a learning objective originating in the algorithmic fairness literature that asks for a predictor to satisfy accuracy-in-expectation simultaneously on many groups. In addition, we include a lower bound showing that multiaccuracy is in fact a necessary property of any non-trivial algorithm that satisfies multigroup robustness.

With these sufficient conditions in hand, we present an efficient post-processing approach that can be used to augment any existing learning algorithm to add both multigroup robustness and multiaccuracy guarantees, while preserving the performance guarantees of the original learning algorithm.

We supplement our theoretical results with experiments on real-world census datasets demonstrating that our post-processing approach can be added to existing learning algorithms to provide multigroup robustness protections without a drop in accuracy.

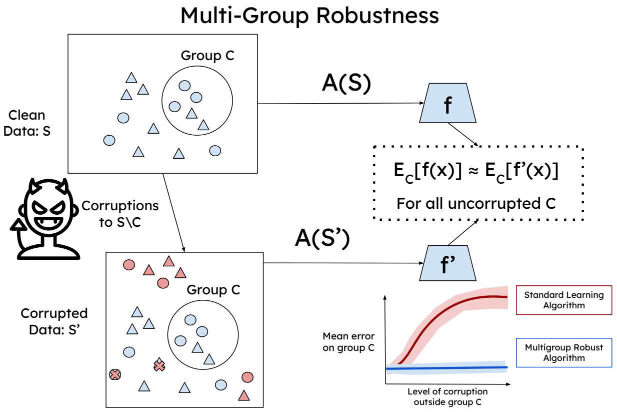

Figure 1: Intuitive illustration of multigroup robustness: for every group C, if points within the group are not modified, a multigroup robust algorithm produces a predictor that achieves marginal mean consistency with the clean data predictor.

[*] Stanford University. Supported by Moses Charikar's and Omer Reingold's Simons Investigators awards and the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Email: lunjia@stanford.edu

[†] Stanford University. Supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Email: cpeale@stanford.edu

[‡] Stanford University. Supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Email: jhshen@stanford.edu