**Author Response: Classification Under Misspecification: Halfspaces, Generalized Linear Models, and Evolvability**

We would like to thank all the reviewers for taking the time to understand the contributions of our paper, and for their helpful comments and/or suggestions. We do not have much to add, but would just like to emphasize a few points in response to the reviews.

We think our contribution relative to the breakthrough work of Diakonikolas et al. is not just that our algorithm is proper or that the insights behind it lead to algorithms for more general concept classes, but that by avoiding partitioning the domain into a polynomial number of regions, it actually becomes practical and something that we can run on real data.

We think that the experimental results are striking, but still only a proof-of-concept in the sense that we added the noise to the data ourselves. A truly compelling demonstration would be, like the first reviewer said, if we could find some real data where our algorithm works well and is demonstrably more fair. This is a direction we are actively pursuing, but we feel that it is a substantial project and would likely be a separate paper if it is successful. Nevertheless some works in fairness work with a graphical model whose causal structure produces confounding effects that lead off-the-shelf algorithms to produce unfair decision rules, see e.g. Kusner et al. [2017]. These types of models naturally lead to situations where noisy observations of some latent quality score might be more variable for some demographics than for others.

Also, we agree that it is hard to do justice to all the technical ingredients in just 8 pages. We attempted to give a more detailed outline for our proper learner, and just some of the key ideas for GLMs. It is an interesting suggestion that we could have split it into two papers. However the results actually build on each other, e.g. our algorithm for GLMs in the $\zeta = 0$ case depends on some knowledge distillation primitives which in turn use our proper learning algorithm for halfspaces.

# References

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.