Thank you for the thoughtful feedback and comments; we are delighted to see this positive response. All reviewers agree that we tackle an important problem of interest to the NeurIPS community, and acknowledge our extensive/insightful experimental analysis which shows improvement over state-of-the-art. The reviewers recognize that we present an interesting idea of relying on computational-identifiability [R3] and our proposed adversarial reweighing method is sound [R3] which is composed in a well-written paper [R1,R3,R4]. Thank you! We respond to your questions below:

[R1] *When are DRO's benefits apparent ...?* The main benefit of DRO is that it induces distributional robustness. Consequently, DRO is guaranteed to focus on worst-case risk for *any* group in the data exceeding size α. In contrast, ARL would only improve the performance for groups that are computationally identifiable groups over (x,y). However, in the fairness setting, this is not necessarily a benefit for DRO, as the DRO authors state in their ICML talk: "it leads to a trade-off between the outlier resistance and fairness, and when the fraction of outliers in training data exceeds size α DRO is completely broken." Our experiments on label bias (Fig. 3(b)) shed some light on this. These experiments were performed with DRO parameter α set to 0.2. We observe that while the fraction of incorrect ground truth class labels (i.e., outliers) is less than 0.2, the performance of ARL and DRO is nearly the same. As the fraction of outliers in the training set exceeds 0.2 we observe that DRO's performance drops significantly. We will revise related work, and experiments section to reflect this more clearly.

[R1] *The employed reweighting approach is standard, while it offers great performance in the considered problem.* While the reweighted optimization problem is standard, the key contribution of this paper is in the adversarial approach of learning these weights over computable groups in (x,y). Our experiments on "ARL vs Inverse Probability Weighting" (Tbl. 2) illustrate the advantages of ARL over standard re-weighting approaches.

[R1] *Is there a unified approach that performs best for all possible scenarios?* Thank you for your thoughtful comment. This along with your other comment on equipping robustness ignited interesting discussions amongst the authors. We think combining the strengths of computationally identifiable regions with distributional robustness would be an interesting research direction to unify the strengths of the two approaches in future research.

[R3] *Regarding smoothness of computationally-identifiable regions, and avoiding overfitting to outliers:* Your observations, and understanding is correct. The design and complexity of the adversary model f_ϕ plays an important role in controlling the granularity (and smoothness) of computationally-identifiable regions of error. More expressive f_ϕ lead to finer-grained upweighting but runs the risk of overfitting to outliers. We currently discuss this on line 200-202, but for further clarity we will expand this discussion, and include experimental results on the relationship between the adversary's complexity and the smoothness of the function learnt by the adversary. Further, recall that the adversary example weight function $\lambda_\phi(x_i, y_i) : f_\phi(x_i, y_i) \to \mathbb{R}$ in Eq. 5 is over $f_\phi(x_i, y_i)$. Hence, the distribution of training weights $\lambda_\phi(x_i, y_i)$ is a good indicator of the smoothness of the function f_ϕ learnt by the adversary, and a practical heuristic to check if the model is overfitting to training outliers.

[R3] *Additional related work:* Thank you for pointing to relevant related work Coston et al. [AIES 2019] and Chen et al. [FAccT 2019]. We will update our related work section. Coston et al. focus on domain adaptation of fairness in settings where the group labels are known for either source or target dataset. Chen et al. investigate the reasons for bias in estimating unfairness in the settings where proxies for group labels are used to evaluate model unfairness.

[R3] *Line 184 defines regions as $l(h(x), y) > \epsilon$. But the loss function in Eq. 5 does not adhere to this.* We intended to use this notation from a descriptive perspective as a framing device to explain what we mean by a "significant error". However, this is not necessary for optimization. We will update the model section to clarify this.

[R4] *COMPAS protected groups are not computationally-identifiable ... surprising:* While we did not perform a formal study on this, we observed that the adversarial predictive accuracy for race in COMPAS is 0.61 is consistent with prior work [1]. We believe that there are demographic signals in the data, but they are not strong enough to predict groups well.

[R4] *How should one understand if computationally-identifiable groups will include the protected groups?* Even evaluating fairness without any group labels is very difficult[2], and as such with *zero* information this would be hard to do. However, we expect in many scenarios practitioners have a variety of domain knowledge they can leverage. For example, if groups labels were not available in training but for a small curated validation set or in an auxiliary dataset, one could evaluate if the computationally identified groups include protected groups, e.g., by analyzing the learnt weights (as in Fig. 4). Practioners might also have knowledge about their domain, e.g., skin tone is present in images and heavily correlated with race. We will add a discussion on limitations of computational identifiability to the paper, and update the broader impact sections to reflect this.

[R4] Thank you for helpful additional feedback. We will address all the comments by appropriately revising the text, and reiterating some of the earlier limitations on label noise in broader impact sections for further clarity.

[1] See Fig. 4 in "iFair: Learning individually fair data representations for algorithmic decision making. In ICDE 2019. "

[2] See "Assessing algorithmic fairness with unobserved protected class using data combination. In FAccT 2020."