We thank the reviewers for providing helpful feedback and for pointing out typos, which we will fix in the final version. **R2** and **R4** note that it would be desirable to have equality of opportunity experiments and code available. While our paper is focused on algorithmically unifying and theoretically clarifying different notions of group fairness and their statistical guarantees, rather than an extensive empirical comparison of different algorithms, we agree these would be good to have and will provide them in the final version of the paper.

Contributions of our framework. First of all, we highlight independent group fairness, which arises naturally from independently requiring fairness for each attribute. This is proposed in existing work e.g. [15], but to our knowledge, we are first to analyze it. A key algorithmic contribution in this case is a plugin approach that keeps track of a confusion matrix for each independent group instead of a confusion matrix for each intersectional group, as an appropriate extension of [20] would have. This is enabled by the inverse group probability weighting (2) and shows that the statistical dependence is *linear* instead of *exponential* in number of groups (which existing analyses suggest). Juxtaposing different notions of fairness also leads one to study relationships between them, e.g., what is the nature of intersectional fairness violations when independent fairness is enforced? We give an example in the appendix. For probabilistic results, we shed light on Bayes optimal predictors in Section 3 and in our plugin algorithm, which are new results. Finally, our framework has the benefit of generality, exactly capturing previous approaches as well as insufficiently explored *multiclass* problems (previously unexplored, and nontrivial to extend to).

Addressing specific comments by reviewers.

- **(R1)** Our framework is flexible and certainly allows one to use OR to define $\mathcal{G}_{\text{fair}}$, or any other grouping dependent on the sensitive attributes (Assumption 2.1); we simply chose the most common cases in practice/previous work (unrestricted, intersectional, independent, gerrymandering) for examples and experiments.
- **(R1)** We emphasize that our paper does focus on *max-violation*, as the fairness problem we state places a constraint on each subgroup. Indeed, as the reviewer notes, the discrepancy between average and max-violation is the source of fairness gerrymandering, and is one of the motivations for this paper. The reason we use the term "average fairness violation" in Proposition 4.1 is that when the Lagrangian is formed for some choice of dual weights, the objective contains an average of the violations; but in the actual algorithm we minimize many different Lagrangians and combine them to produce a classifier which is fair for each group. We will clarify this possible misunderstanding in the final version.
- **(R1)** Thank you for noting issues with understanding notation; which we believe are partially due to a lack of space in combination with the generality we set out to achieve. In the appendix, section C: Estimators walks through explicitly applying the plugin and W-ERM framework to an instance of the fairness problem. This may help, but does not fit in the main paper. We will include a more intuitive description of these methods in the main paper.
- **(R1)** You are right in pointing out that calibration constraints are not linear functions of the confusion matrix; we'll make note of this in the final version.
- **(R1)** The estimation error of eta appears in the middle term in the definition of κ in Theorem 5.1.
- **(R3)** Yes, the converse claim of Proposition 3.2 was that fairness *does not* imply intersectional fairness in general. Remark 3.3 was stating that gerrymandering fairness *does* imply intersectional fairness.
- **(R3)** Theorem 3.1 states that an optimally fair classifier can be constructed from a convex combination of just two weighted classifiers, determined by some weight matrices W_1, W_2. This theorem serves to characterize the ideal solution to our problem.
- **(R3)** You are right: DP should be $C_{0,1} + C_{1,1}$. The $C_{0,0} + C_{1,1}$ is a typo; in our experiments (and in the supplement, section C: Estimators) we used the correct definition. Oops.
- **(R4)** Separating the generalization and optimization claims is ideal; this is essentially what we have done with Theorem 5.1, which isolates the generalization claim for the W-ERM approach. We had to give a separate statement, Theorem 5.2, for the plugin algorithm, because the algorithm involves updating $\boldsymbol{\lambda}$ with empirical violations but minimizing the Lagrangian with respect to the distribution defined by $\hat{\eta}$, which while learned from the empirical distribution is not quite the same. As a result, the proof has components that are hard to untangle.
- **(R4)** Perhaps the discrepancy in times you have noticed is due to an old implementation of W-ERM based on the FairReduction code being used for the Independent experiments but not the Gerrymandering experiments. We will update this table in the final version.
- **(R4)** The reason for the buffer ε is that we would like to compare the classifier we find to the optimally fair classifier, but the optimally fair classifier may violate empirical finite sample fairness constraints.
- **(R4)** The ρ and ρ_g constants are essentially L_1 norms, e.g. for DP all are between 2 and 4.