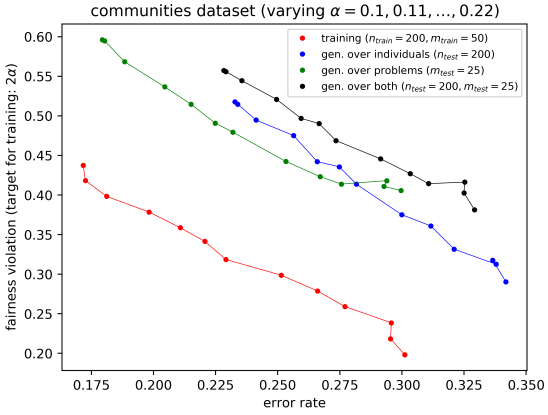
Thanks for the reviews! We take your comments to heart and will make all of the small changes suggested. Here we address the more major concerns.

**Experiments** First, we agree with reviewer 3 that the main contribution of our paper is conceptual and theoretical: in particular, identifying randomness over the problem distribution as something that can be leveraged to give an individual level guarantee, and then deriving the algorithms and generalization bounds needed to actually realize this guarantee. The purpose of the experiments is modest: to support the theory. Because of this, we chose to illustrate experimentally the one thing that is not actually guaranteed by the theory: namely, the convergence of the algorithm. The theorem we prove only guarantees that the algorithm converges *if we have a cost sensitive classification oracle*. In practice, we do not: we use a regression heuristic — and so we chose the experiment we did to confirm that we get convergence, despite the gap between theory and practice. Reviewer 3 is correct that prior work — in particular, [Kearns et al.] — has shown with similar experiments that *other* oracle efficient algorithms converge in practice with heuristics. But that does not imply that ours will, since our algorithm differs both in the specifics of the dynamics (fictitious play in [Kearns et al.], no-regret vs. best response in our paper) and in the particular problems we are asking the classification oracles to solve. In short, because the framework of "oracle efficiency" leaves a gap between theory and practice, we think of it as good practice to accompany a proof of oracle efficiency with an empirical validation of convergence whenever possible.

That being said, Reviewer 3 is of course correct that empirically examining generalization is an interesting thing to do: we didn't do it in the submission simply because we thought it was *less* interesting than convergence (generalization is guaranteed by our theorems, without any heuristic assumptions) and were mindful of space constraints. But we have now quickly adapted our code to investigate our generalization error, both across data points and across problems. The plot is below:




To be consistent with our paper, we trained on exactly the same subset of Communities and Crime that we did in our paper ($n = 200$ datapoints, $m = 50$ problems (selected features from the dataset)). Thus the curve labelled "training" is the same as the reported in-sample results in our paper. We used a fresh holdout consisting of $n = 200$ datapoints, and $m = 25$ problems (features from the dataset that weren't previously used) to evaluate our generalization performance over both problems and data points, in terms of both accuracy and fairness violation. Two things stand out:

1. As predicted by the theory, our test curves track our training curves, but with higher error and unfairness. In particular, the ordering of the models on the Pareto frontier is the same in testing as in training, meaning that the training curve can indeed be used to manage the trade-off out-of-sample as well.
2. The gap in error is substantially smaller than would be predicted by our theory: since our training set is so small, our theoretical guarantees are vacuous, but all points plotted in our test pareto curves are non-trivial in terms of both accuracy and fairness. Presumably the gap in error would narrow on larger training sets. If the paper is accepted and the reviewers feel that generalization experiments should be included, we will make room to do so (and re-run the above experiment on a larger training set, which we can easily do with the luxury of time). Similarly, if the reviewers still feel that the experiments detract from the theory, we are also willing to relegate all experiments to the supplemental material and focus on the main contributions in the body.

**Other Conceptual Questions**

**Reviewer 1:** The group-fairness proposal of [Kearns et al.] indeed mitigates the "gerrymandering" concern we cite from their work, but ultimately does not eliminate it because it remains a group-level constraint that only holds for groups that are sufficiently large. With our approach, we are able to reduce these groups to size 1, which makes it an individual level constraint. (We agree this is different than "individual fairness" in Dwork et al. – we use the term more broadly, and will clarify).

**Reviewer 3:** We agree that asking for parity of error statistics is not always the right approach, and that our techniques generalize to asking for upper bounds on error rates. We consider equalizing error rates mostly as a canonical example of a popular approach; we will clarify. We will also further emphasize the limitation of needing to be able to observe labels for different problems on the same training set.