We thank all the reviewers for their feedback. They are really helpful and will improve the presentation of the paper. We first address the common concerns, and then consider the remaining concerns of each reviewer separately.

- **Experimental Details**: We tried to include most of the important details of the experiment in the main text. We decided to leave out details of the datasets, and hyperparameters from the main text mainly because we view our paper providing a theoretical and algorithmic contribution. However, we realize that the choice of the hyperparameters is important to make our work reproducible, and will include them in the main text in the next version.
- **Hyperparameter Selection**: We first split each dataset into 5 random 80%-20% training and test sets. Then, for the experiments, we split each training set further into a 80%-20% train and validation sets (so there were 5 random sets of 64%-16%-20% train-validation-test). We used the validation sets to select the hyperparameters by performing a grid search over the parameter space. Finally, all the reported accuracies and unfairness gaps (figures 2 and 3) correspond to performance on averages from the 5 original test sets (which were untouched in selecting the hyperparameters). We will update the paper to make the distinction between validation and test sets clear.
- **Definition and Confusion between f and h**: Thanks for pointing out that we are using function f instead of h in lines 87-92. In fact, they are the same and we will change it to h to make this paragraph consistent with other parts of section 2. There is a mistake on the line following eq. (1) and it should read difference in *acceptance* instead of difference of *accuracy*. This means that the definition of demographic parity should just compare the weighted predictions of labels between the two groups and the true label is not a part of the definition. However, the definition of equalized odds do require true labels, as we mention in line 90.

R1: *Re. Validation accuracy and fairness*: validation accuracy and fairness refer to performance on the validation set and they were used to determine the hyperparameters. The reported accuracy and fairness (figures 2 and 3) correspond to performance on the test set, not on the validation set.

Re. Hyper-Parameter Selection: We fixed the number of iterations of algorithm 1 to be 10 for EO and 5 for DP. By theorem 1, increasing this parameter will only increase the accuracy of the final randomized classifier. All the other hyperparameters were chosen by performing grid search using the validation set. The tested values were $\{0.1, 0.2, \ldots, 1\}$ for B, $\{0, 0.05, \ldots, 1\}$ for η, and $\{100, 200, \ldots, 2000\}$ for T.

R2: We agree that it is a nice idea to highlight the technical challenges of the main algorithm, and remove some of the details of the method. This would also allow us to provide more details of the experiment in the main text.

R4: *Re. Weight-Based Characterization of Neighboring Distributions*: Our weight-based characterization can handle sampling biases like under-representation of a group. This can be done by up-weighting the instances from the minority group. In terms of attribute measurement errors, it can handle the situation when errors across all the attributes are similar e.g. they are all under-reported. More general type of attribute errors can be handled by assigning weights to each individual and feature pair and extending our definition for such a set of weighted distributions.

Re. Drop in Accuracy of the Robust Classifier: We also compared the accuracy of the optimal unfair classifier with and without the robustness constraints. We saw a similar drop in accuracy for the ADULT dataset ($\approx 8\%$), and for the COMPAS dataset it was even worse than what we observed for the fair classifier. This explains the drop of accuracy in figure 3 for these two datasets. Moreover, we worked with a large set of weights $\mathcal{W}$ and if we consider a small set (e.g. weights centered around the uniform weight with a small radius), the cost of robust classification will be lower.

Re. Runtime: Across the four datasets, the runtime for a full run of algorithm 1 (on a standard, 2-core 2017 Macbook Pro laptop for $T = 10$ iterations) was 30-40 minutes across all datasets. On the other hand, the standard in-processing based non-robust fair classifier takes about 1 minute. We didn't optimize our code and the runtime can be significantly reduced through multiprocessing (the main bottleneck is solving the linear programs in Algorithm 2).

Re. $h(x, a)$ vs $h(x)$: We need access to the protected groups so that we can verify if a given classifier is unfair or not. Without any access to the protected groups, it is impossible to estimate such biases without making strong assumptions [Kallus et. al. FAT*-20]. In the future, we hope to extend our work for this setting under reasonable assumptions.

R5: *Re. Definition of Demographic Parity*: The line following eq. (1) should read difference in *acceptance* (1-predictions) instead of difference in *accuracy*. This implies that true label is not a part of the definition of DP.

Re. Limited Loss Function: Fairness definitions in binary classification setting are usually considered with 0-1 predictions. This is the reason we considered loss functions over discrete domains. If one defines fairness with real-valued predictions, we can also consider general loss functions. Moreover, for this setting, our algorithms extend immediately – λ-best response is still given by LP, and h-best response still becomes a weighted classification problem.

Re. Paragraph in lines 98-101: We think the concept of randomized classifier and weighted ERM follow from the discussions in the previous paragraphs. We can elaborate the non-convexity of $\mathcal{H}_{\mathcal{W}}$ with an example – if we use logistic regression followed by thresholding, even simple DP fairness constraint makes the parameter space non-convex.