We thank all reviewers for their valuable comments. We first address major comments shared by reviewers and then individual comments.

**Non-asymptotic behavior of the proposed procedure:** In order to present the non-asymptotic result in a clean way we require few modifications to the paper, which we summarize below.

(1) We slightly modify our proposed procedure in Sect. 3, by incorporating part (ii) in Assumption 4.1 directly into the method. Specifically, we explicitly perform the truncation proposed in lines 197–200 with $c_{n,N} = N^{-1/4}$.

(2) Assumption 4.1 now consists only of part (i), since part (ii) has been incorporated in the method.

(3) Finally, we remove Remark 4.6 and modify Theorem 4.5, as follows.

**Theorem 4.5.** *Under Assumptions 2.2 and 4.3, there exist universal constants $C, C' > 0$ such that the proposed algorithm (with truncation) satisfies*

$$\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\Delta(\hat{g},\mathbb{P})] \leq C \sum_{s\in\{0,1\}} \left( \frac{\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{X|S=s}|\eta(X,s)-\hat{\eta}(X,s)|}{\mathbb{P}(Y=1\,|\,S=s)} + \Big(\mathbb{P}(S=s)N\Big)^{-1/4} \right) ,$$

$$\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\mathcal{R}(\hat{g})] \leq \mathcal{R}(g^*) + C' \sum_{s\in\{0,1\}} \left( \frac{\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{X|S=s}|\eta(X,s)-\hat{\eta}(X,s)|}{\mathbb{P}(Y=1\,|\,S=s)} + \Big(\mathbb{P}(S=s)N\Big)^{-1/4} \right) .$$

*Moreover, if the estimator $\hat{\eta}$ satisfies (modified) Assumption 4.1, the proposed algorithm satisfies*

$$\lim_{n,N\to\infty} \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\Delta(\hat{g},\mathbb{P})] = 0 \quad \textit{and} \quad \lim_{n,N\to\infty} \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\mathcal{R}(\hat{g})] \leq \mathcal{R}(g^*) \ .$$

Let us mention that it is possible to write explicit values for the constants $C, C' > 0$, which are independent from the parameters of the problem. We can see that the rate *w.r.t* the size of the unlabeled dataset is $N^{-1/4}$. The rate is non-parametric due to the truncation argument to upper bound the quantity $1/\mathbb{E}_{X|S=s}[\hat{\eta}(X,s)]$. Moreover, let us mention that even in the presence of an infinite number of unlabeled data, the dependence on the $\ell_1$ norm is unavoidable for plug-in methods. Indeed, a close inspection of our proof strategy reveals that the pseudo-estimator (see sketch of the proof of Theorem 4.5) $\tilde{g}$ has this term in its upper bound. Finally, we stress that that in the *classical non-parametric* classification without extra assumptions, the rate of $\ell_1$ norm estimation of $\eta(\cdot)$ is minimax optimal (see [46]) for the classification excess risk.

**Extension to several groups:** We note that the argument in Proposition 2.3 extends to the case that the number of groups $G$ is larger than two. Due to the space limitation, we only sketch the proof using Appendix A as the reference point. In this general case, the constraints read as

$$\mathbb{P}\left(g(X,S)=1\,|\,Y=1,S=s\right) = \mathbb{P}\left(g(X,S)=1\,|\,Y=1,S=s+1\right) \quad \text{for all} \quad s\in\{1,\dots,G-1\} \ .$$

For these constraints it is still possible to write the dual problem introducing $\lambda_1,\dots,\lambda_{G-1}$ real Lagrange multipliers. Similarly to the proof of Proposition 2.3, we first solve the dual formulation which can be performed explicitly. Unlike the case of two groups, which results in one condition on one value $\theta^*$, now we will have $G-1$ different conditions for $G-1$ different values $\theta_1^*,\dots,\theta_{G-1}^*$. Consequently, once the form of the optimal classifier is established, it will be apparent how to extend the plug-in approach to this case following our scheme. However, we feel that this extension is out of the scope of this paper and we prefer to explore it in detail in future work.

**R1.** *"Assumption 2.2 seems, regardless ..."*. We agree with the reviewer that theory and practice might do not always agree with each other. Yet, we care to point out that our theory driven approach shows promising empirical results. In order to remove this assumption one may consider probabilistic classifiers, which is a valuable future research direction.

*"Finally, the experiments left me more puzzled ..."*. Although perhaps we have overstated the good performance of "RF+Ours", please note that, looking at the results in Table 1, "RF+Ours" is among the best performing methods in terms of DEO (since DEO should be as small a possible) except for the "Adult" dataset, where the train/test splits were provided and no cross-validation was performed.

**R2.** *"Isn't it pretty common to assume bounded Rademacher complexity ..."*. It is indeed a common assumption in the study of empirical risk minimizers (ERM). However, there is an important difference between ERM type algorithms and our plug-in approach. The main goal of ERM theory is to approximate the best classifier in a given family of classifiers (*e.g.,* linear classifiers), whereas here we directly aim at estimating the optimal (overall) classifier.

**R3.** *"I'm assuming that $\eta$ is the true underlying probability ..."*. The reviewer is correct. We agree that the phrasing might be misleading; we will modify Assumption 2.2 by avoiding the term "regression function".

*"Experiments: given that one of this work ..."*. We will address in detail the comments by the reviewer in the revised version. During the rebuttal we were able to perform some preliminary experiments on the COMPAS and Adult dataset, which are the only ones big enough to allow performing the requested experiments.

| RF+Ours | COMPAS ACC | COMPAS DEO | Adult ACC | Adult DEO |
|---|---|---|---|---|
| $\mathcal{D}_n=1/10$, $\mathcal{D}_N=1/10$ | 0.68 | 0.07 | 0.79 | 0.06 |
| $\mathcal{D}_n=1/10$, $\mathcal{D}_N=2/10$ | 0.68 | 0.07 | 0.79 | 0.06 |
| $\mathcal{D}_n=1/10$, $\mathcal{D}_N=4/10$ | 0.70 | 0.06 | 0.79 | 0.05 |
| $\mathcal{D}_n=1/10$, $\mathcal{D}_N=8/10$ | 0.71 | 0.05 | 0.80 | 0.04 |

*"...I would like to see more comparisons with direct constrained optimization approaches that work on nonlinear models..."* We care to point out that not only we compared our method with Zafar (which is linear) but also with Donini and Hardt (which works also in the non-linear case). A comparison to Cotter and Agarwal will be inserted as requested in the revised version but we did not manage to do it on time for the rebuttal.

**All.** Finally, we thank all the reviewers for their careful reading. We will address all the minor points (typos, notation issues, figure colors, and remark movements) as underlined and requested by the referees. Of course, we will include the final not anonymous link to the code upon acceptance.