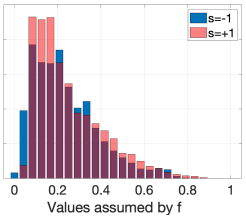
We thank all reviewers for their positive and valuable comments. Let us first address comments shared by multiple reviewers and then answer individual comments.

Relaxation of fairness constraints. In our contribution the smoothing parameter β is introduced to accelerate subgradient methods, however during our experiments we observed that it also controls risk/fairness trade-off. Hence, a relaxation of fairness constraints is implicit in our method via β. It would be interesting to derive statistical theory connected to this parameter in future works. We will add an illustration highlighting this phenomena. Another direction that one can take is to directly relax the fairness constraints asking for $\max_{q\in\mathcal{Q}_L}\mathcal{U}(g,\{q\})\leq\varepsilon$ (see ll. 114–115). Our methodology is flexible enough to allows one to deal with this case and our theory can be adjusted accordingly as well. The price for such relaxation is doubling the dimension of the Lagrange multipliers due to the absolute value in the definition of $\mathcal{U}$. We will sketch the argument in the appendix.

Other notions of fairness. It would indeed be interesting to study extensions of our techniques to other notions of fairness. Though, to the best of our knowledge, there is no general notion of Equalized Odds in regression that would be widely accepted in literature, the main technical difficulty with a possible extension of our machinery to such notion would come from the conditioning on Y.

R1 Visual description. To help the reader we will add the plot to the right, illustrating the distribution of the prediction for L=25 on CRIME.



R2 Discrete regression outputs. We agree that in some practical applications the discrete outputs might be undesirable. First of all note that for sufficiently large discretization parameter L, the discrete nature of the output might be unnoticeable. Secondly, some classical fairness unaware methods such as kNN and decision trees also construct only discrete outputs. Yet, if discrete prediction is still an issue for the problem at hand, a simple ad-hoc remedy is to first fit our method, and then construct an *interpolating* curve (for instance polynomial of sufficiently high degree) using an additional unlabeled dataset. **Errors in multipliers into account.** Let us point out that it is a standard practice in statistical theory to separate the statistical analysis from the numerical optimization problem. Nevertheless, as reviewer pointed out, in order to carry out our proof the dual gradient should be controlled. To this end, note that G_β defined at line 754 is convex and has $2/\beta$-Lipschitz gradient. Thus $\|\nabla G_\beta(\lambda_T)\|_2^2\leq(\beta/4)\{G_\beta(\lambda_T)-\min G_\beta(\lambda)\}$. Furthermore, Thm. D5 in the appendix provides a control on $\{G_\beta(\lambda_T)-\min G_\beta(\lambda)\}$, hence we can control $\|\nabla G_\beta(\lambda_T)\|_2^2$ – the gradient of the smoothed objective. Finally, note that the gradient of $G(\cdot)$ is essentially the argmax function, while the gradient of $G_\beta(\cdot)$ is the soft-arg-max (see Lem. D4). Controlling the deviation of the gradient of $G(\cdot)$ from the gradient of $G_\beta(\cdot)$ would yield the desired bound. Even though this extension is interesting, we feel that it would further complicate already dense proofs with little practical benefits. **Wasserstein fair classification.** Though this work deals with classification and does not tackle the problem of risk minimization under Demographic Parity (DP), we thank the reviewer for this relevant reference. **How do you set slack variable?** All the hyperparameters are tuned according to the scheme described in ll. 249–255. **How does discretization effect violation of constraint?** Actually, the less atoms in discretization, the easier it is to be fair in the sense of DP. Imagine for simplicity that the discretization consists of one point, then the discretized predictors set $\mathcal{G}_L$ (see ll. 113–114) contains only one constant function, which is of course fair. However, if there are too few atoms in the discretization, then the corresponding predictors in $\mathcal{G}_L$ are not powerful enough to yield a good risk guarantee (Lem. 2.4). Thus, the discretization should be balanced between the risk and fairness. **VC-theory.** The intuition of the reviewer is correct. We essentially transform the problem into a *multi-class classification* setup with a specific risk. The main difficulty of course is to understand the amount of classes that one should pick. **Exchange of min and max.** In principle the reviewer is correct that the claim that $\mathcal{R}(\tilde{g}_{\lambda^*})=\mathcal{R}(g^*)$ is equivalent to strong duality. The idea of the proof is to actually establish that this strong duality holds. Note that since $\tilde{g}_{\lambda^*}$ solves the *dual* problem, then $\min\{\mathcal{R}(g)\,:\,g\text{ is fair}\}\geq\mathcal{R}(\tilde{g}_{\lambda^*})$ thanks to the weak duality. To derive the equality note that at line 534 we demonstrate that $\tilde{g}_{\lambda^*}$ is fair. It means that $\tilde{g}_{\lambda^*}$ is feasible for the primal problem, and then $\min\{\mathcal{R}(g)\,:\,g\text{ is fair}\}\leq\mathcal{R}(\tilde{g}_{\lambda^*})$. Thus, $\tilde{g}_{\lambda^*}$ minimizes the primal problem as well. The proof is concluded.

R4 runtime. We note that since our algorithm works in post-processing manner, the most demanding part is the training of the base estimator, which largely depends on the considered algorithm. The runtime of the post-processing Algorithm 1 is present at lines 230-234, which is actually sublinear in the amount of data. In contrast, the algorithm of Agarwal et. al (see their Thm. 1 and Alg. 1) is super linear in the input data and each iteration of their algorithm requires to solve two optimization problems. **Constant C.** This constant is exactly the one appearing in Thm. B6. Unfortunately, we did not find a reference with explicit constant. Our rough computations show that it is at most 36. **The proprietary data [...] what about ethnicity?** The data is predominantly mononational, only less then 1% of students are not of the dominant nation. Due to privacy reasons we cannot disclose the dominant nation of the students.

R5 sacrifices a small amount of error. Note that since we do not minimize the error over *all* predictions, but only over fair ones, the decrease in accuracy is inevitable unless the regression function is fair in the first place. That is, in our case the sacrifice is due to a very effective satisfaction of the fairness constraints.