To all reviewers We would like to thank the reviewers for their thoughtful comments and useful suggestions, which helped us improve the manuscript. Below we provide point-by-point responses.

To Reviewer 1

[R1-1] (Extension to multi-class classifiers) As per your great suggestion, we will present a multi-class classifier counterpart that we developed yet not included in the current draft. The idea is to make a binary hard decision *individually for each element* in the softmax output, and then to estimate the pmf of each hard decision via KDE. For a 3-way classification synthetic dataset (3 Gaussian mixture), this extended method offers respectful improvements over [44] - trends of the gains are similar to those in Figs. 2 and 3. In a revision, we will provide details on our extension together with the experimental results while moving experimental details to the appendix as suggested.

[R1-2] (Hyperparameter tuning) (a) Yes, we exhaustively searched hyperparameters for the baselines. For instance, the learning rate was best-picked among several log-scaled candidates (b) Similarly the narrow MLP was chosen as a result of search - we found the depth and width do not affect too much in performance at least for the considered benchmark datasets. We will clarify these in a revision.

[R1-3] (Motivation of the kernel method and writing-flow) (a) In fact, the stability is also a key issue that we wanted to highlight, although it is not well motivated in the current introduction. We will rewrite the introduction to better balance the issues while citing relevant papers suggested. (b) Yes, the hard/soft decision issue is a key motivation, and the relevant insight allowed us to use the kernel method powerfully and beneficially. For a smooth logical flow, we will re-balance Secs. 3 & 4 so that the issue is strongly emphasized together with an ablation study (w/ and w/o differentiability) in a synthetic setting (as suggested).

[R1-4] (Other comments) Thanks for your suggestion of the "majority"/"minority" naming as well as pointing out typos. We will fix them.

To Reviewer 2

[R2-1] (Lack of theory) We fully agree that the theory is missing for the main claim re. improved performance. Yes, the analysis was not that simple due to the non-convexity of the problem. We will acknowledge this with a proper discussion in a revision.

[R2-2] (Comprehensive experiments for supporting performance gains?) As per your great suggestion, we will conduct an ablation study in which one may be able to separate improvement due to our regularizer from that due to the use of a more flexible model. Specifically we will compare ours to a kernel SVM by Zafar et al. and include this result in a revision.

[R2-3] (Complexity comparison w.r.t. Agarwal et al) As you may guess, the theoretical complexity analysis of our algorithm was not done although we empirically demonstrated that ours exhibits lower complexity relative to Agarwal et al. (requiring multiple rounds of training); see Table 1 in supplementary. Instead we will provide in-depth empirical analysis by plotting the running time as a function of the number of data points, as suggested.

[R2-4] (Problem statement & organization) As per your suggestion, we will make the problem statement clearer and more formal, as well as move the regularizer part into the "Proposed Approach" section.

[R2-5] (Relation to prior work & additional feedback) Thanks for pointing out the approaches (Wasserstein Fair Classifier etc) that directly estimate fairness measures via information theory. We will cite them with enough discussion. Also we will include the kernel SVM by Zafar et al. and Agarwal et al. in the synthetic setting.

To Reviewer 3

[R3-1] (Extension to multi-class settings and performance comparison) As mentioned in response to [R1-1], we actually developed a generalized kernel method that is applicable to multi-class settings. As per your suggestion, we also made performance comparison on a 3-way classification synthetic dataset, observing similar gains as those exhibited in Figs. 2 and 3. We will discuss all of these in a revision.

[R3-2] (Multiple sensitive attributes) Yes, our approach works well for the complex setting. We now conducted experiments on one such setting (AdultCensus with two sensitive attributes: race, gender), observing similar performance improvements, as those in Figs. 2 and 3. We will include the results in a revision.

[R3-3] (Comparison to [44] in many aspects) While our approach offers key benefits in training stability and tradeoff performance, we do agree that [44] is more flexible in terms of application domains. For a fair comparison, we will summarize pros-&-cons of our approach relative to [44] in many aspects.

To Reviewer 4

[R4-1] (Dataset set in Fig. 1 and accuracy of the pmf estimate) We employed a Gaussian mixture: $0.3 \cdot \mathcal{N}(0.37, 0.0055) + 0.7 \cdot \mathcal{N}(0.74, 0.0055)$. Here the true probability is around 0.7, and this is very close to the pmf estimates in Fig. 1(Right). We will provide this in a revision.

[R4-2] (Robustness quantification and its relation with accuracy) One way of quantification is to compute the variance of the pmf estimates over different h's. We will mention this in a revision while plotting accuracy as a function of h.

[R4-3] (Removal of sensitive attributes?) Yes, that is one natural trial. However, such removal does not ensure fairness especially when X is correlated with Z. Please see [42] for details.