We thank the reviewers for the constructive comments. We will revise the paper accordingly.

Reviewer#1

- Our paper exactly deals with the dynamic situation where feature distribution could change over time due to the deployment of classifiers. Regarding the situation where training data arrive sequentially, it is a different setup.

- **Our code is available online** as given in lines 88-89 in the supplementary file.

- As we pointed out in lines 76-79, Dwork's compound decision-making process or pipelines differ from our setting in that individuals drop out at any stage and classification in subsequent stages depends on the remaining cohort of individuals. Wang's paper assumes that multiple functions over the same set of attributes are multiplied to produce an overall score. To the best of our knowledge, our paper is the first work to study the fair learning scenario where there exist multiple related classifiers at different stages and the feature distribution may change due to the deployment of classifiers. We will narrow down our claim to be more accurate.

- For complexity and scalability, if there is only one classifier, then our problem formulation is a convex constrained optimization. If there are multiple classifiers, neither the loss function nor constraints are convex. However, their gradients can be easily computed since each classifier is involved as a single term in the multiplication (e.g., Eqs. (7,8)). Thus, adaptive gradient methods for non-convex optimization such as Adam can be straightforwardly adopted. The convergence of Adam-type algorithms for non-convex optimization has been studied, e.g., in [Chen, et.al. ICLR'19].

- Multiple sensitive attributes are not a bottleneck of the paper as most fairness notions can be easily extended to handle multiple sensitive attributes. For example, $P(y^+|do(s))$ in the total effect can be extended to $P(y^+|do(\mathbf{s}))$ where $\mathbf{s}$ is a value assignment to the combination of multiple sensitive attributes.

Reviewer#3

- The assumption that the causal graph is given is common in the fairness research based on Pearl's structural causal models. In addition to the PC algorithm, there are also quite a number of algorithms to build causal graphs from the data. The sensitivity of causal inference on the learned causal graph structure is beyond the scope of our paper.

- In optimization, we actually add constraints to the objective function as regularization terms. As mentioned in responses to Reviewer#1, the gradients can be easily computed. Then, we adopt Adam for the optimization.

- Regarding complexity, please refer to the corresponding response to Reviewer#1.

- We plan to do experiments with more datasets.

- In this paper we assume Markovian models for simplicity. However, our method can also be extended to scenarios where the Markovian assumption does not hold, a.k.a., semi-Markovian models. As discussed in lines 318-322, we will explore the use of σ-calculus for judging identifiability and computing post-intervention distributions. Furthermore, in the case of unidentifiable, we will resort to bounding approaches to deal with soft interventions.

Reviewer#4

- Clarity. In our paper, we use y_k to denote both classification label and prediction, and use soft intervention to distinguish between them: if the distribution is pre-interventional (i.e., observational), such as $P(y_k^+|\mathbf{z}_k)$, y_k^+ is the label; if the distribution is post-interventional, such as $P(y_k^+|do(\cdots, h_k, \cdots))$ or $P_{h_k}(y_k^+|\mathbf{z}_k)$, y_k^+ is the prediction. After we convert post-intervention distributions to observable distributions, all probabilities are to be estimated from the training data. We originally planed to use y_k to denote the label and $\hat{y}_k$ to denote the prediction. However, this would make the notations too tedious and decrease the readability since most y_k^+, y_k^- in all equations would become $\hat{y}_k^+, \hat{y}_k^-$. On the other hand, from the viewpoint of soft intervention, the prediction is simply an interventional variant of the label upon performing the soft intervention and hence can be distinguished by soft intervention without ambiguity. We will remove notations $\hat{Y}_k$ and $\hat{y}_k$ and more clearly state our notations. Regarding your specific questions: (1) In Definition 1, y_k^+ means the prediction. (2) In lines 211-212, y_i also means the prediction. (3) We will also revise lines 204, 207, and 253-262 of the paper to improve the readability following your suggestions.

- For the separate method, each classifier uses the direct parents of each label and is learned directly from the training data. For the serial method, each classifier also uses the direct parents of each label but is learned from the distribution after upstream classifiers are deployed.

- For the separate method, we did a grid search on τ_1, τ_2 to find classifier pairs h_1, h_2 whose fairness is between -0.05 and 0.05 in training. Then, we evaluated these classifiers in testing and found that in 71.43% of these pairs, h_2 exceeded the interval [-0.05, 0.05] and hence violated the fairness criterion.

In the Adult dataset, `Workclass` and `Income` are two decisions with disproportionate (imbalanced) ratios (31:69 for `Workclass` and 24:86 for `Income`). We oversampled the data twice to adjust the ratios of two decisions to 48:52 and 50:50 respectively which helps us focus on utility-fairness evaluation without distraction from imbalanced classifiers.