We thank all the reviewers for the time devoted to providing thoughtful suggestions. In what follows we respond to points each reviewer made individually.

[Reviewer #1, #4 Comparison with the Theoretical Results in Other Work] Both our work and [4] aim to obfuscate the sensitive attribute using representation learning. However, our theoretical results are different from theirs: Bertran et al. [4] propose the optimization problem where the terms in the objective function are defined in terms of mutual information. Under their formulation, they analyze a trade-off between between utility loss and attribute obfuscation: under the constraint of the attribute obfuscation $I(A;Z) \leq k$, what the maximum utility loss ($I(Y;X \mid Z)$) is. As a comparison, our work provides a trade-off between the 0-1 loss from two groups and attribute obfuscation using attribute inference advantage and a JS-divergence term. Furthermore, we also gave a worst-case guarantee on the attribute inference error that *any* attacker has to incur, which could be used in many real-world applications to give a lower bound estimate of the attack error. We are happy to incorporate the above comparison in our next version of the paper.

[A1] is a preliminary version of [4] and shares the same setting as well as the results as [4], so it is also different from ours.

[Reviewer #2, #3 Extension of our Results] As we have pointed out in our paper, our analysis could be easily extended to the categorical case: For the trade-off between accuracy and attribute obfuscation, we need to first analyze any two values that categorical variable can take on k values using the similar techniques. Then we will have C_k^2 lower bounds, and the sum of the conditional accuracies is lower bounded by the sum of the C_k^2 lower bounds divided by $k-1$. For the formal guarantee, the same proof techniques apply to the categorical case. However, our analysis does not consider the case where A is continuous. We leave this as our future work.

[Reviewer #1, Bounds Too Loose] To the best of our knowledge, our formal guarantee is the first one that applies to all representation learning methods under our context and our bound does not make any assumption of the underlying distribution. This means that the theoretical results are distribution-free and can be applied on any distributions. Given its wide applicability, it is possible that the bound becomes loose on one specific distribution, or a data set. That being said, it is not hard to see that the bound is tight on a degenerate distribution where $A = Y$ almost surely. On top of that, the gap between our bound and the actual error becomes even larger in experiments because we cannot guarantee to obtain the optimal H^* due to the non-convex-none-concave optimization problem when training with neural nets.

[Reviewer #1, Evaluation] We have performed the experiments on how the trade-off parameter λ affects the error rate/utility and shown the results in Figure 1 and Figure 2 (see bars with different λ values). The general trend is: with the increase of λ, both the accuracy of the target task and the ability to predict the sensitive attribute decrease. We will provide more detailed results on this experiment in the next version of our paper.

[Reviewer #3 #4, Minor] With respect to the CE loss in classification: yes, actually that is what we have already used in our experiments. For Lemma 3.1, yes, this lemma will still hold in the case where we have multiple classes in our classification problem. However, it only holds for the CE loss. With respect to the achievability of the optimal h and h_A: yes, when we can obtain the optimal ones, then the original minimax problem reduces to a minimization problem with a regularization term. To the best of our knowledge, in the most general case, i.e., non-convex-non-concave setting, there is no algorithm that can guarantee to converge to (even approximate) Nash-equilibrium. Had we have an algorithm that guarantees to converge to some approximate Nash-equilibria with gap γ, such gap γ could be used in a straightforward way in our Theorem 3.1.

[A1] Bertran, Martin, et al. "Learning data-derived privacy preserving representations from information metrics." (2018).