We thank all the reviewers for the time devoted to provide thoughtful comments.

[Reviewer # 1, Estimating information-theoretic distances] First, we would like to thank the reviewer for the comprehensive and accurate summary of our work. We are happy that the reviewer found our results to be novel and useful. We agree with the reviewer that estimating the distance between the learned representations are intractable, since the sample complexity of estimating Shannon entropy, mutual information and related concepts are exponential in the dimension of the representation space. That being said, one alternative way to do so is to consider the variational representations of f-divergence and use rich parametrized function class (e.g., neural networks) to approximate these distances. For example, recent work [2] on estimating mutual information has empirically shown that such approach often leads to better estimation result than classic approaches based on nonparametric density estimation. From this perspective, the $\mathcal{H}$-divergence in Section 3.2 actually serves as a relaxation of the total variation distance and it equals TVD when $\mathcal{H}$ contains all the measurable functions. Hence the lower bound in Proposition 3.1 gives us a practical way to estimate a proxy of TVD in terms of sum of Type-I and Type-II errors in distinguishing group memberships.

[Reviewer # 1, Total variation in Theorem 3.3] The total variation in Theorem 3.3 is w.r.t the input distributions across groups, i.e., $\mathcal{D}_0(X)$ and $\mathcal{D}_1(X)$. In the remark we use "the distance of representation" to mean "the distance of input distributions". We will clarify this sentence in our final version to avoid such confusion.

[Reviewer # 1, Comparisons with existing work] The results in this paper are distinct from the results in [1]. Specifically, the trade-off given in [1] (Proposition 8) is in terms of the fairness frontier function under the context of cost-sensitive loss. Roughly speaking, it shows that if the two decision functions are dissimilar to each other, the fairness constraint will not harm too much on the target utility. As a comparison, our results (Theorem 3.1 and 3.2) directly give lower bounds on the sum of errors across groups in terms of the difference in base rates as well as the distance of representations. Our results are also different from those in Madras and Zhang et al.: they gave an upper bound on the demographic parity gap in terms of the loss incurred by an adversary (Theorem 5.1), while ours are about lower bounds on the errors of the target task.

[Reviewer # 1, Other questions] We use the notation $\mathcal{P} \ll \mathcal{Q}$ to mean that distribution $\mathcal{P}$ is absolutely continuous w.r.t. distribution $\mathcal{Q}$, i.e., for any measurable event E, if $\mathcal{P}(E) > 0$, then we must have $\mathcal{Q}(E) > 0$ as well. The generator function of KL divergence is indeed $f(t) = t \log t$, and the generator function of the inverse KL divergence is $f(t) = -\log t$. Having identical joint distributions implies that the optimal decision functions are the same across groups, but not the other way around. We also add one more experimental result with $\lambda = 50.0$, and the result is listed as follows. Compared with the existing results in Table 2, we can see a consistent trend.

	$\mathrm{Err}_{\mathcal{D}}$	$\mathrm{Err}_{\mathcal{D}_0} + \mathrm{Err}_{\mathcal{D}_1}$	$\lvert\mathrm{Err}_{\mathcal{D}_0} - \mathrm{Err}_{\mathcal{D}_1}\rvert$	$d_{\mathrm{TV}}(\mathcal{D}_0(\hat{Y}), \mathcal{D}_1(\hat{Y}))$
AdvDebias, $\rho = 50.0$	0.201	0.360	0.112	0.028

[Reviewer # 3] We are happy that the reviewer found our paper to be interesting, theoretically sound and well-written. As stated in the last sentence of the conclusion section, our lower bound naturally implies an algorithm based on instance-reweighting to balance the base rates during fair representation learning. However, the detailed design, analysis and empirical validation of such an algorithm is beyond the scope of current paper. Given that nowadays there are more than tens paper on proposing new algorithms to achieve fairness every year, we believe it would be nice to have a theoretical paper with novel analysis techniques and results to study the fundamental limit of such algorithms. Although it is clear that fairness will compromise utility, before this paper it is still unknown to what extent will it, and how is it related to the difference in terms of base rates across groups. From this perspective, we believe our work is a timely paper that answers the above questions quantitatively. As pointed out by Reviewer 1, our analysis technique using Liese and Vadja lemma is novel and useful. This is of independent interest and we expect its applicability in a broader context.

[Reviewer # 4] We would like to thank Reviewer 4 for the encouraging comments. As explained in Theorem 2.1., Chouldechova and Kleinberg et al. mainly proved that positive rate parity and predictive value parity are in general incompatible. This is an impossibility result between two different notions of fairness. As a comparison, we mainly focus on trade-off between utility and fairness. Furthermore, our Theorem 3.1 is a quantitative result in the sense that it not only gives the impossibility statement when base rates are different, but also gives a lower bound on the error that will be incurred by *any* algorithm. Techniques based on instance-reweighting helps to decrease the difference in base rates, and hence we would expect it to help decrease the lower bound as well. This means that we would incur less drop of utility when learning fair representations. Our current technique does not extend to the definition of equal opportunity, and collecting additional data will not help.

[1]. The Cost of Fairness in Binary Classification. Menon and Williamson, FAT* 2018.

[2]. MINE: Mutual Information Neural Estimation. Belghazi, ICML 2018.