Thanks for the valuable comments! Responses to each reviewer follow.

R1:

> Restrictive definition of fairness

Our results in fact hold for most widely used group-based fairness definitions, including: (i) demographic parity and equality of opportunity (per Lines 120–124), (ii) equalized odds (by adding an additional constraint for $Y = 0$ in Λ_D^{EO}), and (iii) disparate mistreatment, also known as accuracy parity (by using the 0-1 loss and Λ_D^{DP}). Thus, we do not believe our results are narrow in scope.

> Non-binary sensitive features

Our theorem can be generalized to non-binary sensitive features, provided one makes stronger assumptions on the noise (e.g., that it is symmetric). Further developing this is certainly of interest, but studying a binary feature is a common starting point in both the fairness and label noise literature. As ours is the first study of the issue of noisy sensitive features (to our knowledge), we focused on getting a clean and practical result in this important case.

> MC noise model is restrictive

MC learning is an active, widely-used noise model (e.g., [1, 2, 3, 4, 5]). By ensuring that our technique works for this noise model, we have covered the two important and pervasive special cases of CCN [3, 1, 5] and PU learning [2, 4].

> Constraint $\alpha + \beta < 1$ in MC model is restrictive

The constraint imposes no loss of generality: when $\alpha + \beta > 1$, we can simply flip the two labels and apply our theorem. When $\alpha + \beta = 1$, all information about the sensitive attribute is lost. This pathological case is equivalent to not measuring the sensitive attribute at all.

> Unclear/undefined notation

The order of equations in (4) was accidentally swapped (our apologies). For all the other points raised: (i) Λ_D is defined in the immediately following para (Lines 111-115) and examples are given in Equations 2 and 3; (ii) $\bar{L}$ is defined on Line 118, and $\bar{\ell}$ immediately before on Line 117; (iii) regarding redefinition of L_D, in both cases is the same quantity; the RHS in the second usage explicates the involvement of the sensitive attribute.

> (c) "In equation (5) . . . given that α and β don't sum up to 1, what happens to the remaining data points?"

There seems to be a slight misunderstanding: the mixture weights need only sum to 1 *within* and not *across* each corrupted class-conditional. One must take $\Pr(Y_{\mathrm{corr}} = 1)$ into account when reasoning about the samples. As a concrete example, consider a CCN setup where $\Pr(Y = 1) = \frac{1}{2}$, and +ve and –ves have a 0% and 50% chance respectively of having their label flipped. One may verify that in this case, $D_{1,\mathrm{corr}} = \frac{2}{3} \cdot D_1 + \frac{1}{3} \cdot D_0$ and $D_{0,\mathrm{corr}} = D_0$. Clearly, here $\alpha + \beta = \frac{1}{3} \neq 1$. The "missing" weight on D_1 is compensated by there being a greater fraction of corrupted +ves, as one can verify $\Pr(Y_{\mathrm{corr}} = 1) = \frac{3}{4}$. Indeed, the fraction of true +ves remains at $\frac{2}{3} \cdot \frac{3}{4} = \frac{1}{2}$, and so no sample goes missing.

R2:

> Post-processing of Jagielski et al. can be applied to any method, and can handle demographic parity

It is correct that the post-processing method of Jagielski et al. can be applied after any fairness-unaware learner. By contrast, our method can be applied to any *in-process* fairness-preserving learner (e.g. [Donini et al., 2018], [Agarwal et al., 2018], [Zafar et al., 2017b]) In-processing algorithms generally result in better tradeoffs than post-processing (e.g., [Agarwal et al., 2018]). We agree regarding demographic parity, and we will note this in our revision.

R3:

> Suggestions on expanding differential privacy discussion

We appreciate the insightful suggestions, and will expand our discussion accordingly. In particular, we do not foresee any difficulties in combining our approach with any privacy-preserving technique for standard (non-sensitive) features.

References

[1] N. Charoenphakdee, J. Lee, and M. Sugiyama. On symmetric losses for learning from corrupted labels. In *ICML*, 2019.
[2] S. Jain, M. White, and P. Radivojac. Recovering true classifier performance in positive-unlabeled learning. In *AAAI*, 2017.
[3] J. Katz-Samuels, G. Blanchard, and C. Scott. Decontamination of mutual contamination models. *JMLR*, 20(41):1–57, 2019.
[4] R. Kiryo, G. Niu, M. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NIPS*. 2017.
[5] B. van Rooyen and R. C. Williamson. A theory of learning with corrupted labels. *JMLR*, 18(228):1–50, 2018.