We thank the reviewers for the positive feedback. We will answer the questions in the following.

Reviewer1. *Q1: Robustness of the proposed methods against Byzantine workers.*

The robustness of Sign and Median SGD both come from the fact that when performing variable updates, the mean of gradients is replaced by the median of gradients, which is less sensitive to extreme values (i.e., those that are possibly provided by Byzantine workers). As noticed by the reviewer, the perturbation mechanism in Section 4 is gradually converting the estimated statistic from median to mean by adding more noise. Thus, the more noise is added, the less robust the estimated statistic will be. This implies that the robustness of Noisy Sign and Median SGD Byzantine workers depends on the the amount of artificial noise (b in Algorithm 3 and 4). On heterogeneous data with Byzantine workers, the performance of the median-based algorithms can be affected by mainly two factors. One is the gap between median and mean, and the other one is the misleading information provided by Byzantine workers. When the possible effect of Byzantine workers is relatively small (e.g. a small number of Byzantine workers) compared with the median-mean gap, some noise is still preferred to reduce the median-mean gap even though this could amplify the effect of Byzantine workers (there could be a trade-off). For non-heterogeneous data (e.g. iid setting), the median could be very close to the mean and the noise scale b in Noisy Sign and Median SGD should be set small or even 0, the effect of such a small noise might be negligible for both convergence and robustness. We will add a more detailed discussion to the paper.

Reviewer 2. *Q1: What claims can be made in terms of privacy.*

This is a very interesting question. The noise added to the gradient might provide some level of differential privacy while improving the utility of the algorithms in practice. If one knows the upper bound on L_2 norm of all possible gradients, one can use it to calculate the differential privacy cost ϵ using standard privacy accountant (e.g. Abadi et al. [2016]). However, in many cases the L_2 norm bound on gradients is unknown or hard to compute, and gradient clipping is required. It is unclear how the utility of the algorithms will be affected if gradient clipping is applied. We will discuss this question in the paper.

Reviewer 4. *Q1: The increased variance in the treatment.*

Indeed, as mentioned by the reviewer, the perturbation mechanism applied in Noisy Sign and Median SGD can increase variance of the original median estimator while making median closer to mean. This mechanism can be understood as trading bias with variance. One needs to properly deal with the added variance in the optimization algorithms. For example, the optimal learning rate of Noisy Sign and Median SGD can depend on the added variance as implied by Theorem 5 and Theorem 6. In practice, such increased variance can change the best learning rate found by hyperparameter search strategies. We will add more discussion on the effect of increased variance.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.