We would like to thank the five reviewers for their feedback. Upon acceptance, we will include in the final version (a) *improved comparison with prior work*, and (c) *missing references*. We first discuss a few common concerns shared by **Reviewer 1**, **Reviewer 5**, and **Reviewer 6**.

– NUMERICAL EXPERIMENTS: Our experiments in the main paper aim at showing the advantages over DADAM, a decentralized variant of Adam method, developed in [Nazari et. al., 2019]. We recall that the purpose of this paper is to provide both an *algorithmic* and theoretical framework for decentralized variant of adaptive gradient methods (focussing on AMSGrad for illustration purposes). Hence, single-server Adam method does not constitute a baseline for our method, rather its decentralized version does, as plotted in our numerical section. We also highlight the advantages over SGD comparing Figure 2, 3 and Figure 4 in section F of the supplementary material where we note that the proposed algorithm is less sensitive to the learning rate, which is one advantage of adaptive methods. While we acknowledge that the numerical experiments can be improved by adding runs on larger datasets (which we plan on doing for the revised paper), we stress on the fact that the current experiments support our theory. The current experiments we are displaying in our paper are informative on how our newly proposed decentralized framework behaves with respect to baseline methods. In Figure 1 (b), we show a very bad convergence behavior of DADAM on heterogeneous data, in echo of the theoretical divergence that we claim in the paper. Nevertheless, our decentralized framework, using AMSGrad as a prototype, and D-PSGD of [Lian et. al., 2017] are exhibiting great convergence. Our framework is similar and sometimes better than D-PSGD. While D-PSGD is a non-adaptive decentralized method, Figure 1 is convincing on the need for a convergent decentralized adaptive method, thus fixing the divergence issue of DADAM (shown both theoretically and empirically through Figure 1).

Reviewer 1: We thank the reviewer for the comments/remarks on our paper.
– COMPARISON WITH [CHEN ET. AL, 2020]:

– BIAS OF v, THE ESTIMATE OF THE SECOND ORDER MOMENT:

Reviewer 5: We thank the reviewer for valuable comments. We add the following:
– COMPARISON WITH [CHEN ET. AL, 2019] AND [ZHOU, DONGRUO, ET AL., 2018]:

Reviewer 6: We thank you for the valuable comments on our submission. We are revising our paper and will update as soon as it is done. Following is our answer to your questions.

– EXPLANATIONS ON THE ASSUMPTIONS: As rightly mentioned by the reviewer, the stepsize is in order $\alpha_t = 1/\sqrt{T}$. The dependence in d leads to a small learning rate in the presence of large networks but our theorem states that the rate would then be as fast as we present it. Hence, the bound in our Theorem prevails over the intuition that the convergence will be slow due to a small learning rate.

discussion on matrix W when number of nodes is large

Reviewer 8: We thank the reviewer for his/her interest in our paper. Below we address your concerns about our contribution.
– DISCUSSION ON THE MATRIX W: