We thank the five reviewers for their valuable feedback. We first discuss concerns shared by some reviewers:

Reviewer 3, Reviewer 4 and Reviewer 6:

– *On the proof of Theorem 1:* We would like to provide a simpler proof for Theorem 1, using another counter example that satisfies all the assumptions in the paper. Consider a two-node setting with objective function $f(x) = 1/2 \sum_{i=1}^{2} f_i(x)$ and $f_1(x) = \mathbb{1}[|x| \leq 1]2x^2 + \mathbb{1}[|x| > 1](4|x| - 2)$, $f_2(x) = \mathbb{1}[|x-1| \leq 1](x-1)^2 + \mathbb{1}[|x-1| > 1](2|x-1| - 1)$, $W = [0.5, 0.5; 0.5, 0.5]$. The optimal solution is $x^* = 1/3$. Both f_1 and f_2 are smooth and convex with bounded gradient norm 4 and 2, respectively. We also have $L = 4$ (defined in A1). If we initialize with $x_{1,2} = x_{1,2} = -1$ and run DADAM with $\beta_1 = \beta_2 = \beta_3 = 0$ and $\epsilon \leq 1$, we will get $\hat{v}_{1,1} = 16$ and $\hat{v}_{1,2} = 4$. Since we have $|g_{t,1}| \leq 4, |g_{t,2}| \leq 2$ due to bounded gradient, and $\hat{v}_{t,1}$ and $\hat{v}_{t,2}$ are non-decreasing, we have $\hat{v}_{t,1} = 16, \hat{v}_{t,2} = 4, \forall t \geq 1$. Thus, after $t = 1$, DADAM is equivalent to running DGD with a re-scaled f_1 and f_2, *i.e.* running DGD on $f'(x) = \sum_{i=1}^{2} f_i'(x)$ with $f_1'(x) = 0.25 f_1(x)$ and $f_2'(x) = 0.5 f_2(x)$, which has unique optimal $x' = 0.5$. Define $\bar{x}_t = (x_{t,1} + x_{t,2})/2$, then by Th. 2 in [Yuan+, 2016], we have $\alpha < 1/4$, $f'(\bar{x}_t) - f(x') = O(1/(\alpha t))$. Since f' has a unique optima, the above bound implies $\bar{x}_t$ is converging to 0.5 which has non-zero gradient $\nabla f(0.5) = 0.5$.

[Yuan+, 2016] Kun Yuan, Qing Ling, and Wotao Yin. "On the convergence of decentralized gradient descent." – *SIAM Journal on Optimization 26.3 (2016): 1835-1854.*

Reviewer 2 and Reviewer 6:

– *Novelty of the contribution:* The novelty of our design is twofold. First, we aim at bridging the realms of decentralized optimization and adaptive gradient methods. The study of adaptive and decentralized methods are conducted independently in the literature. Second, our gossip technique is not the direct average consensus mechanism used in the extensively studied DGD. We will add more discussion on why the direct average consensus mechanism in Decentralized Gradient Descent cannot be used in our case. The main contribution of this work is the rigorous convergence analysis of adaptive gradient methods in decentralized setting and the proposed convergent algorithm `Decentralized AMSGrad`. To the best of our knowledge, and given the non convergence of DADAM, this is the first success application (with rigorous convergence guarantee) of adaptive methods in decentralized optimization.

Reviewer 1. – *More explanations on notations:* Notations will be explained and simplified in the revised paper.
– *Better presentation of line 41-45:* Lines 41-45 simply highlight that our setting is different from [Reddi et al., 2019], not arguing that their approach is incorrect. We will revise this part to avoid confusion.
– *Assumption A2 is strong:* A2 is necessary for the analysis of adaptive gradient methods and is standard in the literature. In the decentralized literature, this assumption might be viewed as strong since only the convergence of SGD-like algorithms has been dealt with so far. Relaxing A2 is interesting but it is out of the scope of this work.
– *Similar ideas on consensus of step-size:* Thank you for providing the relevant references. [2] averages the predefined stepsizes across iterations to make it more tolerant to staleness in asynchronous updates. [3] does not explicitly apply consensus on stepsize but rather allows the stepsize on each node to be different (the maximum difference depends on the graph structure) for deterministic strongly convex problems. Our learning rate consensus is *across workers* instead of across iterations and we allow the adaptive learning sequence on different nodes to be completely different. Our technique and motivation are thus different from these works. A discussion will be added.

Reviewer 2. – *Connection to counter example in [Reddi et al., 2019]:* Both our example and the one in [Reddi et al., 2019] use the idea that sample dependent learning rate can lead to non-convergence. Yet, in decentralized setting, the sample dependent learning rate is caused by different nodes having different adaptive learning rate sequences, while in [Reddi et al., 2019], the non-convergence is caused by over-adaptivity of the adaptive learning rate of ADAM.

Reviewer 3. – *More discussion on Theorem 2 and Alg. 2/3:* Given the non convergence of DADAM (Alg.2), we develop the decentralized variant of AMSGrad (Alg.3) and establish its non asymptotic analysis in Theorem 2. Alg.3 and its respective finite time analysis (Th.2), proving the convergence of our method, are the main contributions of the paper. Th.2 also gives an insight on the role that the number of nodes N plays in the speed of convergence.
– *Tuning ϵ for different algorithms:* We will include this as a tunable hyperparameter in the future experiments.

Reviewer 4. – *Clarify line 164:* [Nazari et al., 2019] claims that DADAM achieves $O(\sqrt{T})$ regret, but with a non-standard regret for online optimization. We prove that DADAM can fail to converge which contradicts their convergent result. The reason is that the convergence measure defined in [Nazari et al., 2019] may hide this non-convergence issue.
– *A large N leads to high communication cost:* Indeed, there is a trade-off between communication and computation in practice. The optimal N depends on the ratio between the speed of computation and communication.

Reviewer 6. – *Bounded gradient assumption is strong:* This assumption is commonly assumed in the literature of adaptive gradient methods since the analyses for these algorithms are way more complicated than that for SGD. Relaxing this assumption is an interesting question but it will be out of the scope of this paper.
– *Advantages over SGD in numerical experiments:* Our experiments in the main paper aim at showing the advantages over DADAM. The advantages over SGD are highlighted comparing Figure 3 and Figure 4 in Appendix D where we note that the proposed algorithm is less sensitive to the learning rate, which is one advantage of adaptive methods.