We would like to thank four reviewers for their valuable feedback. Please find below the corresponding replies.

Reviewer 1. *Q1: More explanations on notations:* We will improve the the notations and remind the reader more frequently in the paper in the revised paper.

Q2: Better presentation of line 41-45: Line 41-45 simply highlights 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.

Q3: Assumption A2 is strong: Assumption A2 is necessary for the analysis of adaptive gradient methods and it is fairly standard in the literature. In the decentralized literature, this assumption might be viewed as a strong one since only the convergence of SGD-like algorithm has been dealt with so far. Relaxing this assumption is interesting and important but it is out of the scope of this work.

Q4: Similar ideas on consensus of step-size: Thanks for providing the relevant references. [2] averages the predefined stepsize sequence across a few iterations to make it more tolerant to staleness in asynchronous variable updates. [3] does not explicitly apply consensus on stepsize, instead, they allow the stepsize on different nodes to be slight 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 is thus different from these works. We will add a discussion on these works in our next version.

Q5: More experiments: Experiments with larger datasets and complex models are under production.

Reviewer 2. *Q1: Connection to counter example in [Reddi et al., 2019]:* Both our example and the example in [Reddi et al., 2019] use the idea that sample dependent learning rate can lead to non-convergence. The difference is that in decentralized setting, the sample dependent learning rate is caused by the fact that different nodes can have different adaptive learning rate sequences, while in [Reddi et al., 2019], the non-convergence is caused by over-adaptivity of adaptive learning rate of Adam.

Q2: Highlight the novelty of the algorithm design: Thank you for your suggestion. 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 gradient methods and decentralized are conducted independently in the literature. To the best of our knowledge, this is the first success application (with rigorous convergence guarantee) of adaptive gradient methods in decentralized optimization. Second, the gossip technique we use is not the direct average consensus mechanism used in DGD which has been extensively studied. We will add more discussion on why the direct average consensus mechanism in DGD cannot be used in our case.

Reviewer 3. *Q1: More rigorous proof for Theorem 1:* []

Q2: More discussion of Theorem 2 and Algorithm 2/3: We will add more interpretation of the theoretical results and algorithms to improve clarity.

Q3: Tuning ϵ for different algorithms: We will include this as a tunable hyperparameter in future experiments.

Reviewer 4. *Q1: Is Theorem 1 stepsize dependent?:* []

Q2: Clarify line 164: [Nazari et al., 2019] claims that DADAM achieves $O(\sqrt{T})$ regret as an online algorithm, but with a non-standard regret for online optimization. We prove that DADAM can fail to converge which is in some sense contradicting their convergent result. The reason might be that the convergence measure defined in [Nazari et al., 2019] can hide this non-convergence issue.

Q3: A large N leads to high communication cost: Indeed, there will be a trade-off between communication and computation in practice. Discussion on this will be added. The optimal N depends on the ratio between the speed of computation and communication.

Reviewer 6. *Q1: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.

Q2: 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. It can be seen that the proposed algorithm is less sensitive to the change of the learning rate, which is one advantage of adaptive gradient methods.

Q3: Theorem 1 violates bounded gradient assumption: []