Reviewer 1**:**

Main Issues: 1) The idea of the paper could be interesting however the paper loses several points in terms of presentation. In particular, even if the paper provides a notation paragraph, it is still difficult for the reader to understand the notation used in the algorithms and the theorems. It requires a lot of effort from the reader (if it is not familiar with the related work) to understand several theoretical parts. For example what the $g_{t,i}^2$ in the algorithms means (if one is not familiar with adaptive methods the square does not make sense since $g_{t,i}$ is the gradient vector). What $\mu_{t,i}$ means? What is the difference of $u_{t,i}$ and $v_{t,i}$? Are these simply constants or they denote something important for the algorithm? (Step-size etc.)

2) In lines 41-45 the work of Reddi et al. 2020 is presented with goal to provide the motivation and reasoning of proposing more efficient method. However, the presentation is not clear? Why Reddi et al. 2020 is not the correct approach? The authors assume that the reader knows the results from Reddi et al. 2020. More explanation and details are required to motivate the proposed algorithms.

3) The assumptions for the proposed theorems are very strong (especially A2) . See the discussion in [1] where the limitation of these assumptions were discussed.

4) The consensus of step-size is not really novel. Similar ideas have been used before in asynchronous decentralized protocols. See for example the works [2] and [3] below.

5) On numerical experiments: The numerical experiments are not adequate. The experiments focus on showing the benefit of the proposed method compared to DADAM in heterogeneous data (for the homogeneous data all the methods are almost identical). More experiments are required to justify the benefits of the proposed method. Training CNN on MNIST is the minimum requirement. The proposed algorithms need to checked in more challenging problems. See Assran et al. 2019 for some examples of more advanced experiments.

Reviewer 2**:**

Maybe the authors could further highlight the novelty of their algorithm design. For now, it seems that the key to the success of Algorithm 2 is the gossip technique used to ensure the consensus of the variance estimation v^t. This technique is very standard in the decentralized optimization literature which limits the novelty of the paper.

Typos: line 6 of Algorithm 2, please define the operation r_t (or describe its general purpose) line 268, missing a bracket in $O(\sqrt{d}/\sqrt{T})$

Reviewer 3**:**

The primary weakness of the paper is that its theory is not as strong as it claims. For example, while I am convinced that Theorem 1 is true, it is not fully proven in this work. Towards the second half of the proof, the authors use imprecise language such as "and that the unbalanced weights on the two functions yields a different minimizer" that do not allow the reader to immediately fill in the gaps. This proof needs significant revisions to be fully accurate. While this may seem to be a minor point, it is actually critical. The literature on Adam is full of incorrect proofs that have later been noticed and corrected, only for the correction to be wrong. Given this history, it is vital to make sure that future contributions to this topic are sound.

While Theorem 2 seems to be rigorously proved (Disclaimer: I have not verified the full details in the appendix) it still suffers from a lack of clarity. For example, Algorithm 2 defines $\hat{v}$ as a function r( ) of past gradients, yet r( ) is never referenced in the text. The reader must deduce from context that it is some arbitrary function, and that its behavior is linked to the satisfaction of of the conditions outlined in L225-226. In particular, the $O(1/\sqrt{(T)})$ convergence results only holds for certain functions r. I believe this should be discussed explicitly, especially because the conditions in L225-226 are relatively opaque. While the bound on U_t is analogous to a condition in (Chen et al., 2018), as the authors claim, the other condition does not seem to have an exact mirror in that paper, at least as far as I can tell. The fact that it is not obvious to me, as a reviewer, suggests that much more explanation is needed for the layperson reading this paper. As such, I think this is a key aspect that can be improved.

In short, the proofs and discussion of the theorems in question should be revised to be as clear as possible. Currently, they are not in a state where I can recommend them for publication, as they are missing key details.

In short, the clarity of the paper is not at the level where I can fully recommend acceptance. In particular, Theorem 1 should have a stronger, rigorous proof. Similarly, the discussion around Algorithms 2 and 3 should be tightened up so that properties that are discussed in vague terms are made mathematically precise (as an example, I believe that this is executed very well in Chen et al., 2018 in discussing generalized Adam properties).

It would also be useful for the reader if the discussion around Theorem 2 was improved. In particular the relevance of the r( ) function to the conditions in L225-226 that are referred to as "necessary" for convergence. Moreover, if these

conditions are truly necessary, a theorem showing this would improve the content of the work. Given that these changes are somewhat substantial, and should have another review before publication for correctness, I cannot recommend publication at this time.

Reviewer 4**:**

1. Theorem 1 shows that for some problems DADAM doesn't converge. However, the proof looks not that rigorous. Basically, we need to choose a specific step size to guarantee the algorithm's non-convergence. Is it step size - dependent? Is it possible that for other step sizes, DADAM converges? 2. Please clarify line 164 " We conjecture that this inconsistency is due to the 165 definition of the regret in [Nazari et al., 2019]." What is the inconsistency? 3.Line 229-line 231 mentions the larger the number of nodes is, the more stable the training process is. However, does a large N leads to high communication cost. If so, there might be some trade-off here.

Reviewer 6**:**

- Convergence analysis is based on the bounded gradient assumption, which is strong. Many functions are not satisfied this assumption such as a quadratic function.

- Numerical experiments were carried out on a simple deep learning network and a dataset. They does not show an obvious advantage over DGD.

- Proof of Theorem 1: the 1-d optimization problem is an incorrect counter-example for DADAM, since it violates the bounded gradient assumption of Assumption 3 for DADAM. Note that Assumption 2 in this paper also is violated by the optimization problem.

- Given the framework of DADAM and existing consensus algorithms (applied to the consent of adaptive learning rates) , this reviewer considers this work is incremental. It is not too difficult to derive a similar framework to get a converge result to stationary points of the regret function.

- The algorithm in Reddi et al. [2020] should not be called a decentralized method since it needs a server node. For a decentralized framework, every node is a worker, and it only exchanges information with some workers.

- The message in Figure 1 is not clear to me: do authors want to show a better convergence rate for decentralized AMSGrad or a better test accuracy. For the former, the learning rate tuning should not be based on validation accuracy. For the letter, DADAM might converge to a different local minimum.

- The authors should run the counter-example given in Theorem 1 with decentralized AMSGrad to see if it converges.