We first discuss some shared questions by the reviewers.

– **Numerical Experiments (to R1, R5, R6, and R8):** Our experiments serve as a support of our theory that decentralized AMSGrad can converge while DADAM cannot. We recall that the purpose of this paper is to provide both an *algorithmic* and theoretical framework for decentralized variants of adaptive gradient methods. Though our experiments showed some advantages of decentralized AMSGrad over D-PSGD of [Lian et. al., 2017], this is not our main purpose. Figure 1 shows the divergence issue of DADAM is negligible on homogeneous data but can be a big problem on heterogeneous data, highlighting the need for convergent algorithms in practice. Exploring more benefits of the proposed algorithms is indeed an important and interesting question for practitioners. However, to be honest, the authors currently do not have enough resources to scale up the experiments since it requires setting up a distributed computation environment on more machines and common free computation resources do not support it. To the best of our knowledge, adaptive gradient methods are never used in decentralization with rigorous guarantees, we sincerely hope the reviewers can evaluate our contribution base more on our algorithmic framework and theoretical analyses.

– **Discussion on the matrix W (to R6 and R8):** The way to set W is not unique, a common choice for undirected graph is the maximum-degree method in [Boyd et. al. "Fastest mixing Markov chain on a graph.", 2004] (denote d_i as degree of vertex i and $d_{\max} = \max_i d_i$, this method sets $W_{i,i} = 1 - d_i/d_{\max}$, $W_{i,j} = 1/d_{\max}$ if $i \neq j$ and (i,j) is an edge, and $W_{i,j} = 0$ otherwise, a variant is $\gamma I + (1-\gamma)W$ for some $\gamma \in [0,1)$). This W ensures assumption A4 for many common connected graph types. A more refined choice of W coupled with a comprehensive discussion on λ in our Th. 2 can be found in [Boyd et. al. "Fastest mixing Markov chain on graphs with symmetries.", 2009], e.g., $1-\lambda = O(1/N^2)$ for cycle graphs, $1-\lambda = O(1/\log(N))$ for hypercube graphs, $\lambda = 0$ for fully connected graph. Intuitively, λ can be close to 1 for sparse graphs and to 0 for dense graphs. This is consistent with the bound in Th. 2, which is large for λ close to 1 and small for λ close to 0 since average consensus on sparse graphs takes longer.

R1: We thank the reviewer for the remarks.

– **Comparison with [Chen et. al, 2020] ([C20]):** [C20] is one of a few recent attempts to use adaptive gradient methods with the periodic model averaging technique in federated learning. [C20] use the parameter server to maintain a synchronized adaptive learning rate (lr) sequence to ensure convergence, leading to local AMSGrad (LAMS). Our setting is different since *a central server is not available*, thus we use an average consensus mechanism to gradually synchronize adaptive lr. Since both decentralized AMSGrad (DAMS) and LAMS use AMSGrad as the prototype, they reduce to similar ones if local iterations $k = 1$ in LAMS and the graph is fully connected in DAMS. The key difference is we study how to use adaptive gradient methods in decentralized optimization **without** a parameter server, rather than under federated learning settings. As asked by the reviewer, it is indeed possible to extend the periodical averaging technique used in [C20] to our decentralized setting. The resulting algorithm will execute line 7,8,11 every k iterations and $\tilde{u}_{t,i}$ will not be updated in local iterations. We expect our result to have a similar dependency on k as in [C20], i.e., the big-O rate will not be affected for $k \leq O(T^{1/4})$ and applies to our framework.

– **Bias in second moment estimation:** We will not have [mean of square of gradients] vs [square of mean of gradients] issue in most cases. Using the same AdaGrad example with $\hat{v}_{t,i} = \frac{1}{t}\sum_{k=1}^{t} g_{k,i}^2$, when t is large and ϵ is small, the adaptive lr $\tilde{u}_{t,i}$ is close to its tracking target $\frac{1}{N}\sum_{i=1}^{N}\hat{v}_{t,i} = \frac{1}{Nt}\sum_{i=1}^{N}\sum_{k=1}^{t} g_{k,i}^2$, which is the mean of square of stochastic gradients. This estimation is unbiased if we want to estimate second moment of stochastic gradients over the optimization trajectory. It could also be biased if we want to estimate the second moment at recent iterations, as the distribution of stochastic gradients could change with t. The effect of the bias on the training is usually problem-dependent. Killing the bias is possible by drawing fresh samples of stochastic gradients to estimate the adaptive lr with extra computation cost.

R5: We thank you for the valuable comments.

– **Comparison with [Chen et. al, 2019] ([C19]) and [Zhou et al., 2018]] ([Z18]):** We compare Th. 3.1 of [C19] with our Th. 2. The term multiplied by C_1 in our Th. 2 have similar a source as the terms multiplied by C_1 and C_4 in Th. 3.1 of [C19]. The terms multiplied by C_4 and C_5 in our Th. 2 have a similar source as the terms multiplied by C_2 and C_3 in [C19]. The other terms in our Th. 2 are caused by *consensus errors* of variable and adaptive lr. We also compare Th. 5.1 in [Z18] with our Th. 3. The C_1' terms in Th. 3 are counterparts of M_1 and M_3 terms in [Z18], the C_4' term corresponds to the M_2 term in [Z18]. [Z18] can show an extra improved rate assuming sparse gradients. We will add more detailed comparisons in our paper.

R6: Thank you for the thorough analysis.

– **Dimension dependency:** Our stepsize and convergence rate depends on dimension because of assumption A3, which implies the variance of the gradient estimator is linear in d. Under such an assumption, even the best rate of SGD is $O(\sqrt{d}/\sqrt{T})$ with stepsize being $O(1/\sqrt{Td})$, similar to our results. It is possible to improve the dependency on d by assuming the total variance of gradient is independent of d, which will lead to $O(1/\sqrt{T})$ rate with $O(1/\sqrt{T})$ stepsize.

R8: We thank the reviewer for the constructive feedback and interest in our paper. We provide more discussions on the numerical experiments and choice of W on the left page, we hope they could address the reviewer's questions.