We sincerely thank the four reviewers for their valuable feedback. We address the following concern common to **R4**, **R5** and **R6** about the nature of our contribution:

– Purpose of FED-LAMB: In the context of Federated Learning, in particular in the cross-device settings with a large volume of devices, training deep neural networks is a burning challenge. Given the potentially cumbersome amount of data present in each device, being able to learn high dimensional and nonconvex models per device is of utmost importance. – The nature of our contribution is thus *to improve the local optimization method for each device* so that a better local model is learned in fewer iterations, leading to *a natural improvement of the communication efficiency* of the federated method, hence requiring less rounds of communication to reach similar accuracy.

R1: We thank the reviewer for valuable comments. A proofreading is being done as we clarify:

– Notations and Precisions: $\bar{L}$ denotes the sum of the smoothness constants and is stated in the supplementary material. The function $\phi(\cdot)$ is set to the identity function in our runs. The typo in Corollary 1 has been fixed. The usage of d is replaced by i. λ is a weight decaying parameter similar to the original LAMB method. It is tuned on a grid-search for our experiments. This is added to the paper.

– Bounds on Theorem 1: The bounds do depend on the number of devices. We precise that our theoretical results hold when **all** devices are selected. Thus the total number of devices is n, as defined in (1) and appears in our bound, similar dependence is observed in related works.

– Bounds on Corollary 1: The bound does depend on L which is the total number of layers. It also depends on the total smoothness, included in the $\mathcal{O}$ notation. The dependence on the number of devices n is in the denominator of the RHS, which is in accordance with the bound of local AMS in Chen et. al. 2020.

R3: We thank the reviewer for valuable comments. Our point-to-point response is as follows:

– Partial selection of devices: The partial selection of devices has practical virtue which we respected in the numerical experiments. Though, as far as convergence bounds, it is common in the literature to consider the total number of workers participating in each round. Either for simplicity or to avoid cumbersome notations, deriving the result for the general case is not an obstacle for the understanding of the convergence behaviour.

– Theorem 1 for multiple local updates: We agree that the assumption that $T = 1$ is rather simplistic. We managed to derive the result for *multiple local updates* in time for the supplementary deadline. Please refer to Theorem 3 in the supplementary for the desired result.

R4: We thank the reviewer for the analysis. We add:

– Various remarks: We agree that strictly speaking, the LAMB technique we introduce in our federated method is not a modification of Adam as in the original paper. Yet, Line 56, we explicitly state that we develop a variant of local AMSGrad using the same layerwise adaptivity technique. We would argue that Local AMSGrad is used as a backbone and that periodic averaging is used as the most efficient way to compute a global model from several local ones.

– Comparison with "Adaptive Federated Optimization": We thank the reviewer for the reference. The difference is, that paper does SGD updates at local workers with ADAM optimizer applied to the global model, while our method runs layerwise-adaptive AMSGrad at local workers with standard FedAvg for global aggregation. It would be interesting to compare these two different schemes. Thanks for the suggestion.

R5: We thank the reviewer for valuable comments. Our response is as follows:

– Notations: We have added some clarification on several notations in our revised paper. $p_{r,i}^t = \{p_{r,i}^{\ell,t}\}_{\ell=1}^{\mathsf{L}}$ is the ratio computed at round r, local iteration t and for device i.

– Comparison with FedBN: We thank the reviewer for this reference that we did not consider. After careful reading of the contribution, we argue that our method is purely on the optimization algorithm aspect of things while FedBN is supposedly a new modeling consideration. Our method can be used with any model, a large variety for numerical runs and certain model satisfying our assumptions for the theoretical part. Batch Normalization could even be an option on top of Fed-LAMB. We agree that it would be interesting to include a comparison with FedBN in our numerical runs to compare how better or worse such layerwise adaptivity is when performed on the algorithmic level.

R6: We thank the reviewer for valuable comments. We clarify the following points:

Assumption H5: Specializing the upperbound of the estimation of the second order moment is doable and would lead to similar result. For the sake of simplicity, we assumed a constant upperbound.

– Convergence of Fed-LAMB: The purpose of this paper is to improve the communication efficiency of the overall learning method in the federated setting. In other words, our method is better than baseline in the sense that for fewer rounds of communications, it reaches a similar accuracy (in terms of stationary point and objective loss function) than other methods. Fair comparison are thus given when the number of local updates are equal for each method. Indeed, the difference vanished when T goes to infinity but we claim that for a small number of local iterations, which is appealing in practical settings where devices can only compute during a short span of time, our method does better. The bound in (6) does not depend on the number of local iterations T since by assumption $T = 1$. Its extension for multiple local updates is given in the supplementary materials, see Theorem 3.