We sincerely thank the four reviewers for their valuable feedback. We would like to address the following concern common to Reviewers **R4:**, **R5:** and **R6:** about the nature and originality 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 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. Our point-to-point response is as follows:

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** 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 which we included in the $\mathcal{O}$ notation. The dependence on then 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. A proofreading is being done we clarify that:

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.

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 thorough analysis. Our remarks are listed below:

Comparison with "Adaptive Federated Optimization":

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

Notations:

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

$p_{r,i}^t$ is the ratio computed at round r, local iteration t and for device i. $p_{r,i}^{\ell,t}$ denotes its component at layer ℓ

Assumption H5:

Comparison with baselines like SGD and ADAM: We are doing Federated Learning.