We would like to thank the four reviewers for their insightful and constructive feedback.

R1&R2&R3&R5: -Further comparison with other related schemes: We would like to highlight that the main contribution of this work is to improve the convergence analysis of FL algorithms that use sketching. Therefore, our main baseline are algorithms based on sketching which also benefit from privacy property. Yet, per reviewers request we provide more comparison as follows. First, note that there are two main approaches to reduce communication cost that is based on, firstly, local computation or, secondly, gradient compression. We note that the state-of-the-art of FL algorithm regarding reducing communication rounds comes from [1] (without compression). In addition, as pointed out correctly by R3:, since sketching is a special case of gradient compression based algorithm we need to compare with [2]. Yet, we would like to highlight that both number of communication rounds and number of bits per communication rounds corresponding to [1] and [2], **with only focusing on unbiased compression (which does not require extra error correction step but have higher compression error compared to biased compression schemes)**, has been further improved in [3]. As discussed in Section. 3, we improve results in [3] **1) by extending them to unbiased compression using `HEAPRIX` which has similar compression property to biased schemes like** top_k and **2) bi-directional communication property due to lower dimension of sketching and 3) not communicating models, unlike [3]**. The comparison with recent work in [4] which uses bidirectional compression based scheme is provided in Section B in the Appendix. Finally, we note that privacy property is one of the main reason that sketching is used for communication which is not the focus of the aforementioned schemes since they are not using sketching.

R1: We thank the reviewer for valuable comments and references. We would like to make the following clarification:

-Discussion on the assumptions: We note that the dependency between the compression ratio and the convergence rate is not an assumption but it is rather a property induced by using count sketching and all the constants are discussed in detail in comparison with prior methods, Section 4.3. The assumptions we use are standard in the relevant literature. Yet, we will gladly add more clarification to the subsequent version of our paper. **-Notation:** Thanks for the comment. We will move algorithm to the bottom of the page.

R2: We thank the reviewer for the useful comments and typos. Our point-to-point response is as follows:

-Numerical Runs: We present in Section D of the Appendix, additional runs on CIFAR-10 showing similar performance of our method. The number of local updates τ has been set to 1 and 5 in the main text and we added runs with $\tau = 2$ in the Section D of the Appendix as well. Larger number of local updates τ tend to undermine the learning performance as we have observed empirically. In the heterogeneous setting, increasing τ can present a risk of learning bad local models. We acknowledge that there is a trade-off to be found here between speed of convergence and the quality of the local models (to obtain a good global one). **-Typos:** Thanks for mentioning typos and we will fix them. And ℓ indicates randomly selected entries in Algorithm 2, we will fix this in the subsequent version.

R3: We thank the reviewer for valuable comments. We clarify the following point on the comparisons:

Comparison with other compressors: Your comment regarding the reference [2] is valid as sketching is a special case of compression, and we will definitely cite this paper and discuss it in a subsequent version. Besides, we would like to highlight that recently [3] improves the communication cost of [2] focusing on unbiased compressor. When using sketching, since our work extends the performance of unbiased compression schemes to biased compressor **due to the use of `HEAPRIX`, bi-directional compression property of sketching, lower dimension of the communicated models**. As a result, in addition to having privacy property of using sketching, we also improve the communication cost of [2] as well as **removing error feedback framework**. We will include these discussions in a subsequent version.

R5: We thank the reviewer for valuable comments. Below we address your concerns:

Additional Numerical Experiments: Additional runs on CIFAR-10 are presented in the Appendix (Section D). While runs with different ratio of active devices at each iteration are interesting, we reported results with a practical one (half of the devices) for illustrative purposes. We agree that rigorously comparing the number of bits transmitted between FedSGD and our methods is important. Yet, we give the important values of 12 and 75 compressing ratio yielding a good order of magnitude on this latter quantity. Our method being almost as fast as FedSGD, despite the high compressing ratio, shows its benefits.

[1] Karimireddy, Sai Praneeth, et al. "Scaffold: Stochastic controlled averaging for on-device federated learning." arXiv preprint arXiv:1910.06378 (2019).

[2] Basu, Debraj, et al. "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations." Advances in Neural Information Processing Systems. 2019.

[3] Haddadpour, Farzin, et al. "Federated learning with compression: Unified analysis and sharp guarantees." arXiv preprint arXiv:2007.01154 (2020).

[4] Philippenko, Constantin, and Aymeric Dieuleveut. "Artemis: tight convergence guarantees for bidirectional compression in federated learning." arXiv preprint arXiv:2006.14591 (2020).