We thank the four reviewers for their valuable feedback. We address on a point-by-point basis the concerns of each reviewer in the following:

Reviewer 3: Thank you for the comments and suggestions:
– *Notations:* 2) Thanks for pointing out the typo. It should also be indexed by j on the RHS. 4) We referred to Lemma 1 indeed. 5) We use the name PRIVIX in our paper, while it refers to the method in [Li et al. 2019]. 6) The Top-k operation alone is biased which is what we want to avoid with the HEAVYMIX update. 7) We chose to encapsulate the update allowing to compute the sketch of the gradient vector. This allows to use any sketching method as a plug-and-play update. 8) HEAVYMIX does require to extract the exact Top_m values of the gradient. 9) See 7) 10) Our statement here is in a general sense that one can not get exact values of the transformed data. While in [Li et al. 2019], the authors tried to address the privacy formally, there seemed to have some technical problem. In this paper, we do not address this point particularly—yet, it is true that the exact data is hidden by the sketches. We will make this point clearer in the revision. 11) The function $S_j(\cdot)$ is any sketching operator that depends on the random hash tables used. This will be clarified. 12) We will add a table to summarize the metrics from the plots.

Reviewer 6: We appreciate your support and for greatly summarizing our contributions.

Reviewer 7: We will proofread the paper and add the following remarks:

– *Privacy:* Indeed, by privacy we mean that the adversary cannot get the exact data (as opposed to standard Federated Learning), since it is hidden in the random sketches. We will make it more clear in the revision.

– *Compressors and Sketches:* Thank you for the references. There are many possible ways for compressed communication. Compression and quantization are two alternatives to the overall objective of making algorithms communication efficient. Though, our paper focuses on sketching techniques in the spirit of tackling the privacy issue of FL along its communication efficiency bottleneck, with improved convergence rates compared with prior methods.

– *Numerical Experiments:* We stress on the observable gap between our method and FedSGD in the numerical runs. FedSGD is a method using the full gradient at each round of communication and thus displaying a higher computation cost than any other methods using sketches that we plot. While (50×100) may seem large, it still represents and $12\times$ compressing ratio, which is considerable. Under such communication reduction, for (50×100) sketch size, the test accuracy is very close (if not identical) to FedSGD in the bottom two figures in Figure 1 and 2. Thus, we believe our results validates the benefit of the proposed methods in practice.

– *Corollary 1 Conclusion:* The proof simply follows from Theorem 3 in in [Horvath and Richtarik, 2020] and Algorithm 2. As rightly mentioned by the reviewer, by setting $\Delta_1 = \frac{d}{m}$ and $\Delta_2 = 1 + c\frac{d}{m}$ we obtain $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} = c\frac{d}{m} + 1 - c$ for the compression error of `HEAPRIX`. We will add this modification in the revised paper.

Reviewer 8: Thank you for your valuable reviews. We address your concerns below:

– *Numerical Experiments:* We have developed numerical experiments on both MNIST and CIFAR-10 under iid and non-idd settings using shallow and deeper neural networks for completeness. Those runs are also reported in the supplementary material. More experiments are in progress for inclusion in the revised paper. Thank you for the suggestion.

– *Comparison with other unbiased compressors:* The comparison with other sketching techniques, which constitute the main baselines for our methods, is extensively studied in the main paper, see Section 4.1, 4.3 and Section B of the supplementary material where we summarized the comparison between bounds in a table. Regarding compression, we believe that most work using for instance quantization methods are out of the scope of this paper since they do not leverage the privacy-induced property of the sketching operation, which is one of our goal in our paper.

When using sketching, our work extends the performance of unbiased *compression schemes* to biased compressor *due to the use of* `HEAPRIX`*, and benefiting from bi-directional compression property of sketching, via lower dimension of the communicated sketches and not sharing models*. Hence, in addition to having privacy property of using sketching, we also improve the communication cost of [Basu et al., 2019], while *removing error feedback framework*.

Hence, for the sake of the 8 pages limit, we did not include any comparison with such compressors but we will include these discussions in a subsequent version.

– *Heuristic behind FedSKETCH:* The algorithmic design is built upon two previous works. In [Ivkin et al. 2019], only the top-K coordinates (heavy hitters) are recovered, while in [Li et al. 2019], the whole model is compressed without specifically addressing the coordinates with largest magnitude. The HEAPRIX method combines the best of both worlds. Thus, faster convergence is achieved with better empirical performance as displayed in our contribution.

Algorithm 3 can then be used with any other sketching or compression techniques in-lieu of the HEAPRIX operation (Line 5 and 14), yet no guarantees are provided that such resulting algorithm will have the same convergence behaviour. The idea behind our contribution is to both leverage the sketching technique for privacy purpose and the unbiasedness of the operation for convergence purpose.