1 Reviewer 6UvC (5;4):

The studied problem to reduce communication cost in federated learning is interesting and important. The proposed methods in this paper are based on a new sketching technique. Theoretical results on convergence are also provided. Experiments on real datasets show that the proposed methods can outperform the baselines adopted in the paper.

The main shortcoming of this paper is that the experiments are not convincing enough. Firstly, only one simple and small dataset (MNIST) is used for evaluation. Secondly, the number of workers for evaluation is 50, which is small for federated learning settings. Thirdly, the performance is measured in terms of communication rounds, which is not enough. For example, the FedSKETCHGATE is based on the full gradient, and the time for one update in FedSKETCHGATE and FedSKETCH is different. Hence, wall-clock time is a more suitable metric for evaluation. Fourthly, only sketch-based baselines are used for comparison. Other communication-efficient methods, such as gradient sparsification based methods, should also be adopted for comparison to improve the convincingness.

Our reply:

2 Reviewer XPvc (4;4):

Although this work provides a thorough analysis of the convergence rates of distributed SGD under different settings, I feel the novelty is insufficient given that most of the techniques have appeared in related works. In addition, the proposed schemes have limited improvement upon previous results. For instance, the proposed compression scheme reduces the MSE of the previous scheme from to , but when , both schemes have the same asymptotic errors.

Moreover, my major concern is the correctness of the main theorems. In Algorithm 3, every clients sketch their local gradient updates based on the same hash table , so are correlated for different client . Therefore when averaging across clients' updates, the variance may not decrease with

. More specifically, I don't think equation (4) (the second equality on Page 19 of the supplementary material) holds. On the other hand, if every clients use different hash tables, then the server cannot simply average the sketched updates from all clients.

In addition to the above issue, I have the following minor concerns/suggestions:

The authors mention that some of the previous sketch-based methods (such as [17]) may not preserve clients' privacy; however, the authors provide no privacy analysis/guarantees for the proposed schemes. Since both and do not include any privatization step, I don't think they will satisfy differential privacy or other forms of privacy guarantees. Some notations are ambiguous. For instance, in the 11-th line of Algorithm 2, the operator is undefined. (Based on the context that each client uses the same hash table, I believe the authors mean). Also, in the proof of Lemma 1 (on Page 18 of the supplementary material), I believe

what the authors tried to bound is the unsketched graidents, i.e.

instead of

.

The bounded variance statement in Property 1 is very confusing. Is it a high-probability bound or a mean square error bound(i.e. for the "with probability at least " statement, what precisely is the probability with respect to)?

Our reply:

3 Reviewer rTPQ (5;3):

Unfortunately, the presentation is not clear, which makes the paper hard to read and evaluate the contributions. I think this is the biggest weakness of the paper, and here are a few suggestions to perhaps improve the writing:

Probably the biggest improvement could come from making the writing less dense, focussing on the main points, and clearly conveying the key ideas in the approaches. For instance, Section 2.2 and Alg 1. are quite hard to understand. The algorithm description seems unclear, it seems like some indentations are missing which would improve clarity (this is true of latter algorithms as well), and even some steps such as 3 and 4 in the algorithm are not clear (what does it mean to set a variable to be (1 +/- something)? (Minor: calling C the compression operator seems strange, since it maps from $R^d to R^d. There is also a typo in line 96.) Similarly, Section 3 and 4 could also be made easier to read. I would perhaps foc$
$In 3.1 line 154 local and global seem interchanged. For Thm 1 parenthesis should be put appropriately for terms in the$

One of the main desiderata in the FL setting is privacy, and the paper spends a lot of time discussing it. However, there do not appear to be privacy guarantees for the proposed algorithms? Since the output of HEAPRIX is the sum of the outputs of HEAVYMIX and PRIVIX on the residual, why is it private since HEAVYMIX keeps the heavy-hitters as they are? It seems that the total communication required in the strongly-convex case with constant condition number is O(d) (ignoring logarithmic terms). This seems to contradict lower bounds for communication for distributed optimization, for e.g. "Towards Tight Communication Lower Bounds for Distributed Optimisation" shows that $\Omega(pd) communication is necessary, where p is the number of machines. I'm not sure that HEAPRIX is novel, simil$
$sketch for the rest, which is very similar to HEAPRIX.$

Our reply:

4 Reviewer cNsG (4;4):

Cons:

The major novel contribution of this paper is the HEARPIX algorithm which combines HEAVYMIX and PRIVIX to preserve the privacy of unsketch operators. From a paradigm perspective, the contributions are a bit limited.

Though the paper discusses sketch-based compressors, only Count Sketch related techniques are discussed. There a rich literature of sketching and sam-

pling based methods, such as random Gaussian matrix, subsampled randomized Hadamard transform [Lu, Dhillon, Foster and Ungar, NeurIPS013], AMS sketch matrix [Alon, Matias and Szegedy, JCSS999], OSNAP matrix [Nelson and Nguyen, FOCS013]. Standard sampling methods including leverage score sampling [Drineas, Mahoney and Muthukrishnan, SODA006]. The paper would definitely have more contributions if an unsketch algorithm that works for more sketching and sampling techniques is proposed and studied.

The experimental results are promising, however, several questions need to be raised: the performance of proposed algorithms in this paper are superior to FetchSGD. However, FetchSGD is a momentum-based method with error-correction for sketching. Intuitively, it should obtain a faster convergence rate compared to vanilla first-order method. It is possible that FedSKETCH gives a better sketching dimension, but it does not make a lot of sense for it to perform much better than a momentum-based method. The big discrepancy in the training loss with local epoch = 1 under heterogeneous is particularly confusing. Another question is some popular datasets for benchmarking federated learning algorithms such as CIFAR-10, FEMINIST and other language tasks are not used. It is worth noting that FetchSGD [Rothchild et al., ICML020] conducts experiments on CIFAR-10, CIFAR-100, FEMINIST with ResNet and PersonaChat with GPT-2. More experiments can be conducted in order to benchmark algorithms in this paper and give a better overall comparison with FetchSGD.

Our reply:

5 Reviewer Y5Qj (6;3):

Originality and quality: The idea of using sketching to improve communication efficiency is not new. Meanwhile, its technical novelty is relatively limited as the proofs seem to closely follow those that appeared in the earlier works. That being said, the comprehensiveness of the theoretical studies and the empirical advantage over other sketching-based algorithms still make this paper a solid step towards communication-efficient FL. I also appreciate the fact that the authors have conducted a detailed comparison of their theoretical results with prior works in the appendix.

Clarity: Overall this paper is clear.

Significance: This paper places itself as the state-of-the-art algorithm among all sketching-based algorithms for FL. In particular, it is not clear to me how the proposed algorithms compare, both theoretically and empirically, with other non-sketching-based approaches (e.g., those based on gradient sparsification and quantization). This limits the significance and applicability of the proposed methods.

Limitations And Societal Impact: Limitations: It would be helpful if the author could compare their algorithms with other non-sketching-based approaches. Moreover, FedSGD seems to be a relatively weak baseline, especially under heterogeneity. Is it possible to combine FedSKETCH with other federated training

algorithms that are more statistically efficient under heterogeneity, say, FedProx (Li et al 2018) or pFedMe (Dinh et al 2020)?

Our reply: