

Appendix for FedSKETCH: Communication-Efficient Federated Learning via Sketching

The appendix is organized as follows: Section A recalls important notations used throughout the paper and provides the formulation of related algorithms used in the main paper and omitted for the sake of the page limit. We present in Section B of this supplementary file, a through comparison with notable related works. Section C contains the proofs of our results and Section D presents additional numerical runs.

A Notations and Definitions

Notation. Here we denote the count sketch of the vector \mathbf{x} by $\mathbf{S}(\mathbf{x})$ and with an abuse of notation, we indicate the expectation over the randomness of count sketch with $\mathbb{E}_{\mathbf{S}}[\cdot]$. We illustrate the random subset of the devices selected by the central server with \mathcal{K} with size $|\mathcal{K}| = k \leq p$, and we represent the expectation over the device sampling with $\mathbb{E}_{\mathcal{K}}[\cdot]$.

Table 1: Table of Notations

p	\triangleq	Number of devices
k	\triangleq	Number of sampled devices for homogeneous setting
$\mathcal{K}^{(r)}$	\triangleq	Set of sampled devices in communication round r
d	\triangleq	Dimension of the model
τ	\triangleq	Number of local updates
R	\triangleq	Number of communication rounds
B	\triangleq	Size of transmitted bits
$R \times B$	\triangleq	Total communication cost per device
κ	\triangleq	Condition number
ϵ	\triangleq	Target accuracy
μ	\triangleq	PL constant
m	\triangleq	Number of bins of hash tables
$\mathbf{S}(\mathbf{x})$	\triangleq	Count sketch of the vector \mathbf{x}
$\mathbb{U}(\Delta)$	\triangleq	Class of unbiased compressor, see Definition 1

Definition 3 (Polyak-Łojasiewicz). *A function $f(\mathbf{x})$ satisfies the Polyak-Łojasiewicz(PL) condition with constant μ if $\frac{1}{2}\|\nabla f(\mathbf{x})\|_2^2 \geq \mu(f(\mathbf{x}) - f(\mathbf{x}^*))$, $\forall \mathbf{x} \in \mathbb{R}^d$ with \mathbf{x}^* is an optimal solution.*

A.1 Count sketch

In this paper, we exploit the commonly used Count Sketch [7] which is described in Algorithm 5.

Algorithm 5 Count Sketch (CS) [7]

```

1: Inputs:  $\mathbf{x} \in \mathbb{R}^d, t, k, \mathbf{S}_{m \times t}, h_j(1 \leq i \leq t), \text{sign}_j(1 \leq i \leq t)$ 
2: Compress vector  $\mathbf{x} \in \mathbb{R}^d$  into  $\mathbf{S}(\mathbf{x})$ :
3: for  $x_i \in \mathbf{x}$  do
4:   for  $j = 1, \dots, t$  do
5:      $\mathbf{S}[j][h_j(i)] = \mathbf{S}[j-1][h_{j-1}(i)] + \text{sign}_j(i) \cdot x_i$ 
6:   end for
7: end for
8: return  $\mathbf{S}_{m \times t}(\mathbf{x})$ 

```

577 **A.2 PRIVIX method and compression error of HEAPRIX**

578 For the sake of completeness we review PRIVIX algorithm that is also mentioned in [26] as follows:

Algorithm 6 PRIVIX/DiffSketch [26]: Unbiased compressor based on sketching.

```

1: Inputs:  $x \in \mathbb{R}^d, t, m, \mathbf{S}_{m \times t}, h_j(1 \leq i \leq t), \text{sign}_j(1 \leq i \leq t)$ 
2: Query  $\tilde{x} \in \mathbb{R}^d$  from  $\mathbf{S}(x)$ :
3: for  $i = 1, \dots, d$  do
4:    $\tilde{x}[i] = \text{Median}\{\text{sign}_j(i) \cdot \mathbf{S}[j][h_j(i)] : 1 \leq j \leq t\}$ 
5: end for
6: Output:  $\tilde{x}$ 

```

579 Regarding the compression error of sketching we restate the following Corollary from the main body
580 of this paper:

581 **Corollary 2.** *Based on Theorem 3 of [15] and using Algorithm 2, we have $C(x) \in \mathbb{U}(c \frac{d}{m})$. This*
582 *shows that unlike PRIVIX (Algorithm 6) the compression noise can be made as small as possible*
583 *using large size of hash table.*

584 *Proof.* The proof simply follows from Theorem 3 in [15] and Algorithm 2 by setting $\Delta_1 = c \frac{d}{m}$
585 and $\Delta_2 = 1 + c \frac{d}{m}$ we obtain $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} = c \frac{d}{m} = O\left(\frac{d}{m}\right)$ for the compression error of
586 HEAPRIX. □

587 B Summary of comparison of our results with prior works

588 For the purpose of further clarification, we summarize the comparison of our results with related
 589 works. We recall that p is the number of devices, d is the dimension of the model, κ is the condition
 590 number, ϵ is the target accuracy, R is the number of communication rounds, and τ is the number of
 591 local updates. We start with the homogeneous setting comparison. Comparison of our results and
 592 existing ones for homogeneous and heterogeneous setting are given respectively Table 2 and Table 3.

Table 2: Comparison of results with compression and periodic averaging in the homogeneous setting. UG and PP stand for Unbounded Gradient and Privacy Property respectively.

Reference	PL/Strongly Convex	UG	PP
Ivkin et al. [17]	$R = O\left(\max\left(\frac{d}{m\sqrt{\epsilon}}, \frac{1}{\epsilon}\right)\right), \tau = 1, B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ $pRB = O\left(\frac{pd}{m\epsilon} \log\left(\frac{d}{\delta\sqrt{\epsilon}} \max\left(\frac{d}{m}, \frac{1}{\sqrt{\epsilon}}\right)\right)\right)$	✗	✗
Theorem 1	$R = O\left(\kappa\left(\frac{d-m}{mk} + 1\right) \log\left(\frac{1}{\epsilon}\right)\right), \tau = O\left(\frac{d}{k\left(\frac{d}{k} + m\right)\epsilon}\right), B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ $kRB = O\left(m\kappa(d-m+mk) \log\left(\frac{1}{\epsilon}\right) \log\left(\frac{\kappa(d\frac{d-m}{mk} + d) \log\left(\frac{1}{\epsilon}\right)}{\delta}\right)\right)$	✓	✓

Table 3: Comparison of results with compression and periodic averaging in the heterogeneous setting. UG and PP stand for Unbounded Gradient and Privacy Property respectively.

Reference	non-convex	General Convex	UG	PP
Basu et al. [3] (with $\gamma = m/d$)	$R = O\left(\frac{d}{m\epsilon^{1.5}}\right)$ $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$ $B = O(d)$ $RB = O\left(\frac{d^2}{m\epsilon^{1.5}}\right)$	—	✗	✗
Li et al. [26]	—	$R = O\left(\frac{d}{m\epsilon^2}\right)$ $\tau = 1$ $B = O\left(m \log\left(\frac{d^2}{m\epsilon^2\delta}\right)\right)$	✗	✓
Rothchild et al. [37]	$R = O\left(\max\left(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon}\right)\right)$ $\tau = 1$ $B = O\left(m \log\left(\frac{d}{\delta} \max\left(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon}\right)\right)\right)$ $RB = O\left(m \max\left(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon}\right) \log\left(\frac{d}{\delta} \max\left(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon}\right)\right)\right)$	—	✗	✗
Rothchild et al. [37]	$R = O\left(\frac{\max(I^{2/3}, 2-\alpha)}{\epsilon^3}\right)$ $\tau = 1$ $B = O\left(\frac{m}{\alpha} \log\left(\frac{d \max(I^{2/3}, 2-\alpha)}{\epsilon^3\delta}\right)\right)$ $RB = O\left(\frac{m \max(I^{2/3}, 2-\alpha)}{\epsilon^3\alpha} \log\left(\frac{d \max(I^{2/3}, 2-\alpha)}{\epsilon^3\delta}\right)\right)$	—	✗	✗
Theorem 2	$R = O\left(\frac{d}{m\epsilon}\right)$ $\tau = O\left(\frac{1}{p\epsilon}\right)$ $B = O\left(m \log\left(\frac{d^2}{m\epsilon\delta}\right)\right)$ $RB = O\left(\frac{d}{\epsilon} \log\left(\frac{d^2}{m\epsilon\delta} \log\left(\frac{1}{\epsilon}\right)\right)\right)$	$R = O\left(\frac{d}{m\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{p\epsilon^2}\right)$ $B = O\left(m \log\left(\frac{d^2}{m\epsilon\delta}\right)\right)$	✓	✓

593 **Comparison with [13] and [36]** Convergence analysis of algorithms in [13] relies on unbiased com-
 594 pression, while in this paper our FL algorithm based on HEAPRIX enjoys from unbiased compression
 595 with equivalent biased compression variance. Moreover, we highlight that the convergence analysis
 596 of FedCOMGATE is based on the extra assumption of boundedness of the difference between the
 597 average of compressed vectors and compressed averages of vectors. However, we do not need this
 598 extra assumption as it is satisfied naturally due to linearity of sketching. Finally, as pointed out
 599 in Remark 2, our algorithms enjoy from a bidirectional compression property, unlike FedCOMGATE
 600 in general. Furthermore, since results in [13] improve the communication complexity of FedPAQ
 601 algorithm, developed in [36], hence FedSKETCH and FedSKETCHGATE improves the communication
 602 complexity obtained in [36].

603 **Comparison with [3].** We note that the algorithm in [3] uses a composed compression and quantiza-
604 tion while our algorithm is solely based on compression. So, in order to compare with algorithms
605 in [3] we only consider Qsparse-local-SGD with compression and we let compression factor $\gamma = \frac{m}{d}$
606 (to compare with the same compression ratio induced with sketch size of mt). For strongly convex
607 objective in Qsparse-local-SGD to achieve convergence error of ϵ they require $R = O\left(\kappa \frac{d}{m\sqrt{\epsilon}}\right)$ and
608 $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$, which is improved to $R = O\left(\frac{\kappa d}{m} \log(1/\epsilon)\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$ for PL objectives.
609 Similarly, for non-convex objective [3] requires $R = O\left(\frac{d}{m\epsilon^{1.5}}\right)$ and $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$, which is
610 improved to $R = O\left(\frac{d}{m\epsilon}\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$. We note that we reduce communication rounds at the
611 cost of increasing number of local updates (which scales down with number of devices, p). Addi-
612 tionally, we highlight that our FedSKETCHGATE exploits the gradient tracking idea to deal with data
613 heterogeneity, while algorithms in [3] does not develop such mechanism and may suffer from poor
614 convergence in heterogeneous setting. We also note that setting $\tau = 1$ and using top_m compressor,
615 the QSPARSE-local-SGD algorithm becomes similar to distributed SGD with sketching as they both
616 use the error feedback framework to improve the compression variance. Finally, since the average of
617 sparse vectors may not be sparse in general the number of transmitted bits from server to devices in
618 QSPARSE-Local-SGD in [3] may not be sparse in general ($B = O(d)$), however our algorithms enjoy
619 from bidirectional compression properly due to lower dimension and linearity properties of sketch-
620 ing ($B = O(m \log(\frac{Rd}{\delta}))$). Therefore, the total number of bits per device for strongly convex and
621 non-convex objective is improved respectively from $RB = O\left(\kappa \frac{d^2}{m\sqrt{\epsilon}}\right)$ and $RB = O\left(\frac{d^2}{m\epsilon^{1.5}}\right)$
622 in [3] to $RB = O\left(\kappa d \log(\frac{\kappa d^2}{m\delta} \log(\frac{1}{\epsilon})) \log(1/\epsilon)\right) = O\left(\kappa d \max\left(\log(\frac{\kappa d^2}{m\delta}), \log^2(1/\epsilon)\right)\right)$ and
623 $RB = O\left(\log(\frac{d^2}{m\epsilon\delta}) \frac{d}{\epsilon}\right)$.

624 Additionally, as we noted using sketching for transmission implies two way communication from
625 master to devices and vice versa. Therefore, in order to show efficacy of our algorithm we compare
626 our convergence analysis with the obtained rates in the following related work:

627 **Comparison with [35].** The reference [35] considers two-way compression from parameter server to
628 devices and vice versa. They provide the convergence rate of $R = O\left(\frac{\omega^{\text{Up}} \omega^{\text{Down}}}{\epsilon^2}\right)$ for strongly-objective
629 functions where ω^{Up} and ω^{Down} are uplink and downlink's compression noise (specializing to our
630 case for the sake of comparison $\omega^{\text{Up}} = \omega^{\text{Down}} = \theta(d)$) for general heterogeneous data distribution.
631 In contrast, while our algorithms are using bidirectional compression due to use of sketching for
632 communication, our convergence rate for strongly-convex objective is $R = O(\kappa \mu^2 d \log(\frac{1}{\epsilon}))$ with
633 probability $1 - \delta$.

634 We would like to also mention that there prior studies such as [43] and [49] that analyze the two-way
635 compression, but since [35] is the state-of-the-art on this topic we only compared our results with
636 these papers.

637 C Theoretical Proofs

638 We will use the following fact (which is also used in [30, 12]) in proving results.

639 **Fact 3** ([30, 12]). *Let $\{x_i\}_{i=1}^p$ denote any fixed deterministic sequence. We sample a multiset \mathcal{P} (with*
 640 *size K) uniformly at random where x_j is sampled with probability q_j for $1 \leq j \leq p$ with replacement.*
 641 *Let $\mathcal{P} = \{i_1, \dots, i_K\} \subset [p]$ (some i_j s may have the same value). Then*

$$\mathbb{E}_{\mathcal{P}} \left[\sum_{i \in \mathcal{P}} x_i \right] = \mathbb{E}_{\mathcal{P}} \left[\sum_{k=1}^K x_{i_k} \right] = K \mathbb{E}_{\mathcal{P}} [x_{i_k}] = K \left[\sum_{j=1}^p q_j x_j \right] \quad (2)$$

642 For the sake of the simplicity, we review an assumption for the quantization/compression, that
 643 naturally holds for PRIVIX and HEAPRIX.

644 **Assumption 4** ([13]). *The output of the compression operator $Q(\mathbf{x})$ is an unbiased estimator of*
 645 *its input \mathbf{x} , and its variance grows with the squared of the squared of ℓ_2 -norm of its argument, i.e.,*
 646 $\mathbb{E}[Q(\mathbf{x})] = \mathbf{x}$ *and* $\mathbb{E}[\|Q(\mathbf{x}) - \mathbf{x}\|^2] \leq \omega \|\mathbf{x}\|^2$.

647 We note that the sketching PRIVIX and HEAPRIX, satisfy Assumption 4 with $\omega = c \frac{d}{m}$ and $\omega =$
 648 $c \frac{d}{m} - 1$ respectively with probability $1 - \frac{\delta}{R}$ per communication round. Therefore, all the results in
 649 Theorem 1, by taking union over the all probabilities of each communication rounds, are concluded
 650 with probability $1 - \delta$ by plugging $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively into the corresponding
 651 convergence bounds.

652 C.1 Proof of Theorem 1

653 In this section, we study the convergence properties of our FedSKETCH method presented in Algo-
 654 rithm 3. Before developing the proofs for FedSKETCH in the homogeneous setting, we first mention
 655 the following intermediate lemmas.

656 **Lemma 1.** *Using unbiased compression and under Assumption 2, we have the following bound:*

$$\mathbb{E}_{\mathcal{K}} \left[\mathbb{E}_{\mathbf{S}, \xi^{(r)}} \left[\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\|^2 \right] \right] = \mathbb{E}_{\xi^{(r)}} \mathbb{E}_{\mathbf{S}} \left[\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\|^2 \right] \leq \tau \left(\frac{\omega}{k} + 1 \right) \sum_{j=1}^m q_j \left[\sum_{c=0}^{\tau-1} \|\mathbf{g}_j^{(c,r)}\|^2 + \sigma^2 \right] \quad (3)$$

Proof.

$$\begin{aligned} & \mathbb{E}_{\xi^{(r)} | \mathbf{w}^{(r)}} \mathbb{E}_{\mathcal{K}} \left[\mathbb{E}_{\mathbf{S}} \left[\left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right\|^2 \right] \right] \\ &= \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\mathbb{E}_{\mathbf{S}} \left[\left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \underbrace{\mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right)}_{\tilde{\mathbf{g}}_{\mathbf{S},j}^{(r)}} \right\|^2 \right] \right] \right] \\ &\stackrel{\textcircled{1}}{=} \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\left[\left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_{\mathbf{S},j}^{(r)} - \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbb{E}_{\mathbf{S}} [\tilde{\mathbf{g}}_{\mathbf{S},j}^{(r)}] \right\|^2 + \left\| \mathbb{E}_{\mathbf{S}} \left[\frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_{\mathbf{S},j}^{(r)} \right] \right\|^2 \right] \right] \right] \\ &\stackrel{\textcircled{2}}{=} \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\mathbb{E}_{\mathbf{S}} \left[\left\| \frac{1}{k} \left[\sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_{\mathbf{S},j}^{(r)} - \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right] \right\|^2 + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right\|^2 \right] \right] \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\left[\text{Var}_{\mathbf{S}} \left[\frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_{\mathbf{S}_j}^{(r)} \right] + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right\|^2 \right] \right] \right] \\
&= \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\left[\frac{1}{k^2} \sum_{j \in \mathcal{K}} \text{Var}_{\mathbf{S}_j} \left[\tilde{\mathbf{g}}_{\mathbf{S}_j}^{(r)} \right] + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right\|^2 \right] \right] \right] \\
&\leq \mathbb{E}_{\xi^{(r)}} \left[\mathbb{E}_{\mathcal{K}} \left[\left[\frac{1}{k^2} \sum_{j \in \mathcal{K}} \omega \left\| \tilde{\mathbf{g}}_j^{(r)} \right\|^2 + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right\|^2 \right] \right] \right] \\
&= \left[\mathbb{E}_{\xi} \left[\left[\frac{1}{k} \sum_{j \in \mathcal{K}} \omega \left\| \tilde{\mathbf{g}}_j^{(r)} \right\|^2 + \mathbb{E}_{\mathcal{K}} \mathbb{E}_{\xi^{(r)}} \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right\|^2 \right] \right] \right] \\
&= \left[\mathbb{E}_{\xi} \left[\left[\frac{\omega}{k} \sum_{j=1}^p q_j \left\| \tilde{\mathbf{g}}_j^{(r)} \right\|^2 + \mathbb{E}_{\mathcal{K}} \left[\text{Var} \left(\frac{1}{k} \sum_{j \in \mathcal{K}} \tilde{\mathbf{g}}_j^{(r)} \right) + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{g}_j^{(r)} \right\|^2 \right] \right] \right] \right] \\
&= \frac{\omega}{k} \sum_{j=1}^p q_j \mathbb{E}_{\xi} \left\| \tilde{\mathbf{g}}_j^{(r)} \right\|^2 + \mathbb{E}_{\mathcal{K}} \left[\frac{1}{k^2} \sum_{j \in \mathcal{K}} \text{Var} \left(\tilde{\mathbf{g}}_j^{(r)} \right) + \left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{g}_j^{(r)} \right\|^2 \right] \\
&\leq \frac{\omega}{k} \sum_{j=1}^p q_j \mathbb{E}_{\xi} \left\| \tilde{\mathbf{g}}_j^{(r)} \right\|^2 + \mathbb{E}_{\mathcal{K}} \left[\frac{1}{k^2} \sum_{j \in \mathcal{K}} \tau \sigma^2 + \frac{1}{k} \sum_{j \in \mathcal{K}} \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] \\
&= \frac{\omega}{k} \sum_{j=1}^p q_j \left[\text{Var} \left(\tilde{\mathbf{g}}_j^{(r)} \right) + \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] + \left[\frac{\tau \sigma^2}{k} + \sum_{j=1}^p q_j \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] \\
&\leq \frac{\omega}{k} \sum_{j=1}^p q_j \left[\tau \sigma^2 + \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] + \left[\frac{\tau \sigma^2}{k} + \sum_{j=1}^p q_j \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] \\
&= (\omega + 1) \frac{\tau \sigma^2}{k} + \left(\frac{\omega}{k} + 1 \right) \left[\sum_{j=1}^p q_j \left\| \mathbf{g}_j^{(r)} \right\|^2 \right] \tag{4}
\end{aligned}$$

657 where ① holds due to $\mathbb{E} \left[\left\| \mathbf{x} \right\|^2 \right] = \text{Var}[\mathbf{x}] + \left\| \mathbb{E}[\mathbf{x}] \right\|^2$, ② is due to $\mathbb{E}_{\mathbf{S}} \left[\frac{1}{p} \sum_{j=1}^p \tilde{\mathbf{g}}_{\mathbf{S}_j}^{(r)} \right] = \frac{1}{p} \sum_{j=1}^m \tilde{\mathbf{g}}_j^{(r)}$.

658 Next we show that from Assumptions 3, we have

$$\mathbb{E}_{\xi^{(r)}} \left[\left\| \tilde{\mathbf{g}}_j^{(r)} - \mathbf{g}_j^{(r)} \right\|^2 \right] \leq \tau \sigma^2 \tag{5}$$

659 To do so, note that

$$\begin{aligned}
\text{Var} \left(\tilde{\mathbf{g}}_j^{(r)} \right) &= \mathbb{E}_{\xi^{(r)}} \left[\left\| \tilde{\mathbf{g}}_j^{(r)} - \mathbf{g}_j^{(r)} \right\|^2 \right] \stackrel{\text{①}}{=} \mathbb{E}_{\xi^{(r)}} \left[\left\| \sum_{c=0}^{\tau-1} \left[\tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)} \right] \right\|^2 \right] = \text{Var} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \\
&\stackrel{\text{②}}{=} \sum_{c=0}^{\tau-1} \text{Var} \left(\tilde{\mathbf{g}}_j^{(c,r)} \right) \\
&= \sum_{c=0}^{\tau-1} \mathbb{E} \left[\left\| \tilde{\mathbf{g}}_j^{(c,r)} - \mathbf{g}_j^{(c,r)} \right\|^2 \right] \\
&\stackrel{\text{③}}{\leq} \tau \sigma^2 \tag{6}
\end{aligned}$$

660 where in ① we use the definition of $\tilde{\mathbf{g}}_j^{(r)}$ and $\mathbf{g}_j^{(r)}$, in ② we use the fact that mini-batches are chosen
661 in i.i.d. manner at each local machine, and ③ immediately follows from Assumptions 2.

662 Replacing $\mathbb{E}_{\xi^{(r)}} [\|\tilde{\mathbf{g}}_j^{(r)} - \mathbf{g}_j^{(r)}\|^2]$ in (4) by its upper bound in (5) implies that

$$\mathbb{E}_{\xi^{(r)}|\mathbf{w}^{(r)}} \mathbb{E}_{\mathbf{S}, \mathcal{K}} \left[\left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right\|^2 \right] \leq (\omega + 1) \frac{\tau \sigma^2}{k} + \left(\frac{\omega}{k} + 1 \right) \sum_{j=1}^p q_j \|\mathbf{g}_j^{(r)}\|^2 \quad (7)$$

663 Further note that we have

$$\left\| \mathbf{g}_j^{(r)} \right\|^2 = \left\| \sum_{c=0}^{\tau-1} \mathbf{g}_j^{(c,r)} \right\|^2 \leq \tau \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|^2 \quad (8)$$

664 where the last inequality is due to $\left\| \sum_{j=1}^n \mathbf{a}_i \right\|^2 \leq n \sum_{j=1}^n \|\mathbf{a}_i\|^2$, which together with (7) leads to
 665 the following bound:

$$\mathbb{E}_{\xi^{(r)}|\mathbf{w}^{(r)}} \mathbb{E}_{\mathbf{S}} \left[\left\| \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right\|^2 \right] \leq (\omega + 1) \frac{\tau \sigma^2}{k} + \tau \left(\frac{\omega}{k} + 1 \right) \sum_{j=1}^p q_j \|\mathbf{g}_j^{(c,r)}\|^2, \quad (9)$$

666 and the proof is complete. \square

667 **Lemma 2.** Under Assumption 1, and according to the FedCOM algorithm the expected inner product
 668 between stochastic gradient and full batch gradient can be bounded with:

$$-\mathbb{E}_{\xi, \mathbf{S}, \mathcal{K}} \left[\left\langle \nabla f(\mathbf{w}^{(r)}), \tilde{\mathbf{g}}^{(r)} \right\rangle \right] \leq \frac{1}{2} \eta \frac{1}{m} \sum_{j=1}^m \sum_{c=0}^{\tau-1} \left[-\|\nabla f(\mathbf{w}^{(r)})\|_2^2 - \|\nabla f(\mathbf{w}_j^{(c,r)})\|_2^2 + L^2 \|\mathbf{w}^{(r)} - \mathbf{w}_j^{(c,r)}\|_2^2 \right] \quad (10)$$

669 *Proof.* We have:

$$\begin{aligned} & -\mathbb{E}_{\{\xi_1^{(t)}, \dots, \xi_m^{(t)} | \mathbf{w}_1^{(t)}, \dots, \mathbf{w}_m^{(t)}\}} \mathbb{E}_{\mathbf{S}, \mathcal{K}} \left[\left\langle \nabla f(\mathbf{w}^{(r)}), \tilde{\mathbf{g}}_{\mathbf{S}, \mathcal{K}}^{(r)} \right\rangle \right] \\ &= -\mathbb{E}_{\{\xi_1^{(t)}, \dots, \xi_m^{(t)} | \mathbf{w}_1^{(t)}, \dots, \mathbf{w}_m^{(t)}\}} \left[\left\langle \nabla f(\mathbf{w}^{(r)}), \eta \sum_{j \in \mathcal{K}} q_j \sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right\rangle \right] \\ &= -\left\langle \nabla f(\mathbf{w}^{(r)}), \eta \sum_{j=1}^m q_j \sum_{c=0}^{\tau-1} \mathbb{E}_{\xi, \mathbf{S}} \left[\tilde{\mathbf{g}}_{j, \mathbf{S}}^{(c,r)} \right] \right\rangle \\ &= -\eta \sum_{c=0}^{\tau-1} \sum_{j=1}^m q_j \left\langle \nabla f(\mathbf{w}^{(r)}), \mathbf{g}_j^{(c,r)} \right\rangle \\ &\stackrel{\textcircled{1}}{=} \frac{1}{2} \eta \sum_{c=0}^{\tau-1} \sum_{j=1}^m q_j \left[-\|\nabla f(\mathbf{w}^{(r)})\|_2^2 - \|\nabla f(\mathbf{w}_j^{(c,r)})\|_2^2 + \|\nabla f(\mathbf{w}^{(r)}) - \nabla f(\mathbf{w}_j^{(c,r)})\|_2^2 \right] \\ &\stackrel{\textcircled{2}}{\leq} \frac{1}{2} \eta \sum_{c=0}^{\tau-1} \sum_{j=1}^m q_j \left[-\|\nabla f(\mathbf{w}^{(r)})\|_2^2 - \|\nabla f(\mathbf{w}_j^{(c,r)})\|_2^2 + L^2 \|\mathbf{w}^{(r)} - \mathbf{w}_j^{(c,r)}\|_2^2 \right] \end{aligned} \quad (11)$$

670 where ① is due to $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2$, and ② follows from Assumption 1. \square

671 The following lemma bounds the distance of local solutions from global solution at r th communication
 672 round.

673 **Lemma 3.** Under Assumptions 2 we have:

$$\mathbb{E} \left[\|\mathbf{w}^{(r)} - \mathbf{w}_j^{(c,r)}\|_2^2 \right] \leq \eta^2 \tau \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \eta^2 \tau \sigma^2$$

674 *Proof.* Note that

$$\begin{aligned}
\mathbb{E} \left[\left\| \mathbf{w}^{(r)} - \mathbf{w}_j^{(c,r)} \right\|_2^2 \right] &= \mathbb{E} \left[\left\| \mathbf{w}^{(r)} - \left(\mathbf{w}^{(r)} - \eta \sum_{k=0}^c \tilde{\mathbf{g}}_j^{(k,r)} \right) \right\|_2^2 \right] \\
&= \mathbb{E} \left[\left\| \eta \sum_{k=0}^c \tilde{\mathbf{g}}_j^{(k,r)} \right\|_2^2 \right] \\
&\stackrel{\textcircled{1}}{=} \mathbb{E} \left[\left\| \eta \sum_{k=0}^c \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \mathbb{E} \left[\left\| \eta \sum_{k=0}^c \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\stackrel{\textcircled{2}}{=} \eta^2 \sum_{k=0}^c \mathbb{E} \left[\left\| \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + (c+1) \eta^2 \sum_{k=0}^c \mathbb{E} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\leq \eta^2 \sum_{k=0}^{\tau-1} \mathbb{E} \left[\left\| \left(\tilde{\mathbf{g}}_j^{(k,r)} - \mathbf{g}_j^{(k,r)} \right) \right\|_2^2 \right] + \tau \eta^2 \sum_{k=0}^{\tau-1} \mathbb{E} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&\stackrel{\textcircled{3}}{\leq} \eta^2 \sum_{k=0}^{\tau-1} \sigma^2 + \tau \eta^2 \sum_{k=0}^{\tau-1} \mathbb{E} \left[\left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \right] \\
&= \eta^2 \tau \sigma^2 + \eta^2 \sum_{k=0}^{\tau-1} \tau \left\| \mathbf{g}_j^{(k,r)} \right\|_2^2 \tag{12}
\end{aligned}$$

675 where $\textcircled{1}$ comes from $\mathbb{E}[\mathbf{x}^2] = \text{Var}[\mathbf{x}] + [\mathbb{E}[\mathbf{x}]]^2$ and $\textcircled{2}$ holds because $\text{Var}\left(\sum_{j=1}^n \mathbf{x}_j\right) =$
676 $\sum_{j=1}^n \text{Var}(\mathbf{x}_j)$ for i.i.d. vectors \mathbf{x}_i (and i.i.d. assumption comes from i.i.d. sampling), and fi-
677 nally $\textcircled{3}$ follows from Assumption 2. \square

678 C.1.1 Main result for the non-convex setting

679 Now we are ready to present our result for the homogeneous setting. We first state and prove the
680 result for the general non-convex objectives.

681 **Theorem 4** (non-convex). *For FedSKETCH(τ, η, γ), for all $0 \leq t \leq R\tau - 1$, under Assumptions 1*
682 *to 2, if the learning rate satisfies*

$$1 \geq \tau^2 L^2 \eta^2 + \left(\frac{\omega}{k} + 1 \right) \eta \gamma L \tau \tag{13}$$

683 *and all local model parameters are initialized at the same point $\mathbf{w}^{(0)}$, then the average-squared*
684 *gradient after τ iterations is bounded as follows:*

$$\frac{1}{R} \sum_{r=0}^{R-1} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 \leq \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)}))}{\eta \gamma \tau R} + \frac{L \eta \gamma (\omega + 1)}{k} \sigma^2 + L^2 \eta^2 \tau \sigma^2, \tag{14}$$

685 *where $\mathbf{w}^{(*)}$ is the global optimal solution with function value $f(\mathbf{w}^{(*)})$.*

686 *Proof.* Before proceeding with the proof of Theorem 4, we would like to highlight that

$$\mathbf{w}^{(r)} - \mathbf{w}_j^{(\tau,r)} = \eta \sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)}. \tag{15}$$

687 From the updating rule of Algorithm 3 we have

$$\mathbf{w}^{(r+1)} = \mathbf{w}^{(r)} - \gamma \eta \left(\frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{s} \left(\sum_{c=0, r}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right) = \mathbf{w}^{(r)} - \gamma \left[\frac{\eta}{k} \sum_{j \in \mathcal{K}} \mathbf{s} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right].$$

In what follows, we use the following notation to denote the stochastic gradient used to update the global model at r th communication round

$$\tilde{\mathbf{g}}_{\mathbf{S},\mathcal{K}}^{(r)} \triangleq \frac{\eta}{p} \sum_{j=1}^p \mathbf{S} \left(\frac{\mathbf{w}^{(r)} - \mathbf{w}_j^{(\tau,r)}}{\eta} \right) = \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right).$$

688 and notice that $\mathbf{w}^{(r)} = \mathbf{w}^{(r-1)} - \gamma \tilde{\mathbf{g}}^{(r)}$.

689 Then using the unbiased estimation property of sketching we have:

$$\mathbb{E}_{\mathbf{S}} [\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}] = \frac{1}{k} \sum_{j \in \mathcal{K}} \left[-\eta \mathbb{E}_{\mathbf{S}} \left[\mathbf{S} \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right] \right] = \frac{1}{k} \sum_{j \in \mathcal{K}} \left[-\eta \left(\sum_{c=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(c,r)} \right) \right] \triangleq \tilde{\mathbf{g}}_{\mathbf{S},\mathcal{K}}^{(r)}.$$

690 From the L -smoothness gradient assumption on global objective, by using $\tilde{\mathbf{g}}^{(r)}$ in inequality (15) we
691 have:

$$f(\mathbf{w}^{(r+1)}) - f(\mathbf{w}^{(r)}) \leq -\gamma \langle \nabla f(\mathbf{w}^{(r)}), \tilde{\mathbf{g}}^{(r)} \rangle + \frac{\gamma^2 L}{2} \|\tilde{\mathbf{g}}^{(r)}\|^2 \quad (16)$$

692 By taking expectation on both sides of above inequality over sampling, we get:

$$\begin{aligned} \mathbb{E} \left[\mathbb{E}_{\mathbf{S}} \left[f(\mathbf{w}^{(r+1)}) - f(\mathbf{w}^{(r)}) \right] \right] &\leq -\gamma \mathbb{E} \left[\mathbb{E}_{\mathbf{S}} \left[\langle \nabla f(\mathbf{w}^{(r)}), \tilde{\mathbf{g}}_{\mathbf{S}}^{(r)} \rangle \right] \right] + \frac{\gamma^2 L}{2} \mathbb{E} \left[\mathbb{E}_{\mathbf{S}} \|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\|^2 \right] \\ &\stackrel{(a)}{=} \underbrace{-\gamma \mathbb{E} \left[\langle \nabla f(\mathbf{w}^{(r)}), \tilde{\mathbf{g}}^{(r)} \rangle \right]}_{\text{(I)}} + \underbrace{\frac{\gamma^2 L}{2} \mathbb{E} \left[\mathbb{E}_{\mathbf{S}} \|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\|^2 \right]}_{\text{(II)}}. \end{aligned} \quad (17)$$

693 We proceed to use Lemma 1, Lemma 2, and Lemma 3, to bound terms (I) and (II) in right hand side
694 of (17), which gives

$$\begin{aligned} &\mathbb{E} \left[\mathbb{E}_{\mathbf{S}} \left[f(\mathbf{w}^{(r+1)}) - f(\mathbf{w}^{(r)}) \right] \right] \\ &\leq \gamma \frac{1}{2} \eta \sum_{j=1}^p q_j \sum_{c=0}^{\tau-1} \left[-\left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 - \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + L^2 \eta^2 \sum_{c=0}^{\tau-1} \left[\tau \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \sigma^2 \right] \right] \\ &\quad + \frac{\gamma^2 L (\frac{\omega}{k} + 1)}{2} \left[\eta^2 \tau \sum_{j=1}^p q_j \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 \right] + \frac{\gamma^2 \eta^2 L (\omega + 1) \tau \sigma^2}{2k} \\ &\stackrel{\textcircled{1}}{\leq} \frac{\gamma \eta}{2} \sum_{j=1}^p q_j \sum_{c=0}^{\tau-1} \left[-\left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 - \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \tau L^2 \eta^2 \left[\tau \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \sigma^2 \right] \right] \\ &\quad + \frac{\gamma^2 L (\frac{\omega}{k} + 1)}{2} \left[\eta^2 \tau \sum_{j=1}^p q_j \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 \right] + \frac{\gamma^2 \eta^2 L (\omega + 1) \tau \sigma^2}{2k} \\ &= -\eta \gamma \frac{\tau}{2} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 \\ &\quad - \left(1 - \tau L^2 \eta^2 \tau - \left(\frac{\omega}{k} + 1 \right) \eta \gamma L \tau \right) \frac{\eta \gamma}{2} \sum_{j=1}^p q_j \sum_{c=0}^{\tau-1} \left\| \mathbf{g}_j^{(c,r)} \right\|_2^2 + \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + \gamma (\omega + 1)) \sigma^2 \\ &\stackrel{\textcircled{2}}{\leq} -\eta \gamma \frac{\tau}{2} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 + \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + \gamma (\omega + 1)) \sigma^2, \end{aligned} \quad (18)$$

695 where in $\textcircled{1}$ we incorporate outer summation $\sum_{c=0}^{\tau-1}$, and $\textcircled{2}$ follows from condition

$$1 \geq \tau L^2 \eta^2 \tau + \left(\frac{\omega}{k} + 1 \right) \eta \gamma L \tau.$$

696 Summing up for all R communication rounds and rearranging the terms gives:

$$\frac{1}{R} \sum_{r=0}^{R-1} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 \leq \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)}))}{\eta \gamma \tau R} + \frac{L \eta \gamma (\omega + 1)}{k} \sigma^2 + L^2 \eta^2 \tau \sigma^2.$$

697 From the above inequality, is it easy to see that in order to achieve a linear speed up, we need to have
 698 $\eta\gamma = O\left(\frac{\sqrt{k}}{\sqrt{R\tau}}\right)$. \square

699 **Corollary 3** (Linear speed up). *In (14) for the choice of $\eta\gamma = O\left(\frac{1}{L}\sqrt{\frac{k}{R\tau(\omega+1)}}\right)$, and $\gamma \geq k$ the
 700 convergence rate reduces to:*

$$\frac{1}{R} \sum_{r=0}^{R-1} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 \leq O \left(\frac{L\sqrt{(\omega+1)} (f(\mathbf{w}^{(0)}) - f(\mathbf{w}^*))}{\sqrt{kR\tau}} + \frac{(\sqrt{(\omega+1)})^2 \sigma^2}{\sqrt{kR\tau}} + \frac{k\sigma^2}{R\gamma^2} \right). \quad (19)$$

701 *Note that according to (19), if we pick a fixed constant value for γ , in order to achieve an ϵ -accurate
 702 solution, $R = O\left(\frac{1}{\epsilon}\right)$ communication rounds and $\tau = O\left(\frac{\omega+1}{k\epsilon}\right)$ local updates are necessary. We
 703 also highlight that (19) also allows us to choose $R = O\left(\frac{\omega+1}{\epsilon}\right)$ and $\tau = O\left(\frac{1}{k\epsilon}\right)$ to get the same
 704 convergence rate.*

705 **Remark 3.** *Condition in (13) can be rewritten as*

$$\begin{aligned} \eta &\leq \frac{-\gamma L\tau \left(\frac{\omega}{k} + 1\right) + \sqrt{\gamma^2 \left(L\tau \left(\frac{\omega}{k} + 1\right)\right)^2 + 4L^2\tau^2}}{2L^2\tau^2} \\ &= \frac{-\gamma L\tau \left(\frac{\omega}{k} + 1\right) + L\tau \sqrt{\left(\frac{\omega}{k} + 1\right)^2 \gamma^2 + 4}}{2L^2\tau^2} \\ &= \frac{\sqrt{\left(\frac{\omega}{k} + 1\right)^2 \gamma^2 + 4} - \left(\frac{\omega}{k} + 1\right) \gamma}{2L\tau}. \end{aligned} \quad (20)$$

706 *So based on (20), if we set $\eta = O\left(\frac{1}{L\gamma} \sqrt{\frac{k}{R\tau(\omega+1)}}\right)$, it implies that:*

$$R \geq \frac{\tau k}{(\omega+1)\gamma^2 \left(\sqrt{\left(\frac{\omega}{k} + 1\right)^2 \gamma^2 + 4} - \left(\frac{\omega}{k} + 1\right) \gamma \right)^2}. \quad (21)$$

707 *We note that $\gamma^2 \left(\sqrt{\left(\frac{\omega}{k} + 1\right)^2 \gamma^2 + 4} - \left(\frac{\omega}{k} + 1\right) \gamma \right)^2 = \Theta(1) \leq 5$ therefore even for $\gamma \geq m$ we
 708 need to have*

$$R \geq \frac{\tau k}{5(\omega+1)} = O\left(\frac{\tau k}{(\omega+1)}\right). \quad (22)$$

709 *Therefore, for the choice of $\tau = O\left(\frac{\omega+1}{k\epsilon}\right)$, due to condition in (22), we need to have $R = O\left(\frac{1}{\epsilon}\right)$.
 710 Similarly, we can have $R = O\left(\frac{\omega+1}{\epsilon}\right)$ and $\tau = O\left(\frac{1}{k\epsilon}\right)$.*

711 **Corollary 4** (Special case, $\gamma = 1$). *By letting $\gamma = 1$, $\omega = 0$ and $k = p$ the convergence rate in (14)
 712 reduces to*

$$\frac{1}{R} \sum_{r=0}^{R-1} \left\| \nabla f(\mathbf{w}^{(r)}) \right\|_2^2 \leq \frac{2(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^*))}{\eta R\tau} + \frac{L\eta}{p} \sigma^2 + L^2 \eta^2 \tau \sigma^2,$$

713 *which matches the rate obtained in [44]. In this case the communication complexity and the number
 714 of local updates become*

$$R = O\left(\frac{p}{\epsilon}\right), \quad \tau = O\left(\frac{1}{\epsilon}\right),$$

715 *which simply implies that in this special case the convergence rate of our algorithm reduces to the
 716 rate obtained in [44], which indicates the tightness of our analysis.*

717 **C.1.2 Main result for the PL/Strongly convex setting**

718 We now turn to stating the convergence rate for the homogeneous setting under PL condition which
 719 naturally leads to the same rate for strongly convex functions.

720 **Theorem 5** (PL or strongly convex). *For FedSKETCH(τ, η, γ), for all $0 \leq t \leq R\tau - 1$, under
 721 Assumptions 1 to 2 and 3, if the learning rate satisfies*

$$1 \geq \tau^2 L^2 \eta^2 + \left(\frac{\omega}{k} + 1\right) \eta \gamma L \tau$$

722 and if the all the models are initialized with $\mathbf{w}^{(0)}$ we obtain:

$$\mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] \leq (1 - \eta \gamma \mu \tau)^R \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\mu} \left[\frac{1}{2} L^2 \tau \eta^2 \sigma^2 + (1 + \omega) \frac{\gamma \eta L \sigma^2}{2k}\right]$$

723 *Proof.* From (18) under condition:

$$1 \geq \tau L^2 \eta^2 \tau + \left(\frac{\omega}{k} + 1\right) \eta \gamma L \tau$$

724 we obtain:

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{w}^{(r+1)}) - f(\mathbf{w}^{(r)})\right] &\leq -\eta \gamma \frac{\tau}{2} \left\|\nabla f(\mathbf{w}^{(r)})\right\|_2^2 + \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + \gamma(\omega + 1)) \sigma^2 \\ &\leq -\eta \mu \gamma \tau \left(f(\mathbf{w}^{(r)}) - f(\mathbf{w}^{(*)})\right) + \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + \gamma(\omega + 1)) \sigma^2 \end{aligned} \quad (23)$$

725 which leads to the following bound:

$$\mathbb{E}\left[f(\mathbf{w}^{(r+1)}) - f(\mathbf{w}^{(*)})\right] \leq (1 - \eta \mu \gamma \tau) \left[f(\mathbf{w}^{(r)}) - f(\mathbf{w}^{(*)})\right] + \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + (\omega + 1) \gamma) \sigma^2$$

726 By setting $\Delta = 1 - \eta \mu \gamma \tau$ we obtain the following bound:

$$\begin{aligned} &\mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] \\ &\leq \Delta^R \left[f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right] + \frac{1 - \Delta^R}{1 - \Delta} \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + (\omega + 1) \gamma) \sigma^2 \\ &\leq \Delta^R \left[f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right] + \frac{1}{1 - \Delta} \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + (\omega + 1) \gamma) \sigma^2 \\ &= (1 - \eta \mu \gamma \tau)^R \left[f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right] + \frac{1}{\eta \mu \gamma \tau} \frac{L \tau \gamma \eta^2}{2k} (k L \tau \eta + (\omega + 1) \gamma) \sigma^2 \end{aligned} \quad (24)$$

727 □

728 **Corollary 5.** *If we let $\eta \gamma \mu \tau \leq \frac{1}{2}$, $\eta = \frac{1}{2L(\frac{\omega}{k} + 1)\tau \gamma}$ and $\kappa = \frac{L}{\mu}$ the convergence error in Theorem 5,*

729 *with $\gamma \geq k$ results in:*

$$\begin{aligned} &\mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] \\ &\leq e^{-\eta \gamma \mu \tau R} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\mu} \left[\frac{1}{2} \tau L^2 \eta^2 \sigma^2 + (1 + \omega) \frac{\gamma \eta L \sigma^2}{2k}\right] \\ &\leq e^{-\frac{R}{2(\frac{\omega}{k} + 1)\kappa}} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\mu} \left[\frac{1}{2} L^2 \frac{\tau \sigma^2}{L^2 (\frac{\omega}{k} + 1)^2 \gamma^2 \tau^2} + \frac{(1 + \omega) L \sigma^2}{2 (\frac{\omega}{k} + 1) L \tau k}\right] \\ &= O\left(e^{-\frac{R}{2(\frac{\omega}{k} + 1)\kappa}} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{\sigma^2}{(\frac{\omega}{k} + 1)^2 \gamma^2 \mu \tau} + \frac{(\omega + 1) \sigma^2}{\mu (\frac{\omega}{k} + 1) \tau k}\right) \\ &= O\left(e^{-\frac{R}{2(\frac{\omega}{k} + 1)\kappa}} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{\sigma^2}{\gamma^2 \mu \tau} + \frac{(\omega + 1) \sigma^2}{\mu (\frac{\omega}{k} + 1) \tau k}\right) \end{aligned} \quad (25)$$

730 which indicates that to achieve an error of ϵ , we need to have $R = O\left(\left(\frac{\omega}{k} + 1\right) \kappa \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau =$
 731 $\frac{(\omega+1)}{k\left(\frac{\omega}{k}+1\right)\epsilon}$. Additionally, we note that if $\gamma \rightarrow \infty$, yet $R = O\left(\left(\frac{\omega}{k} + 1\right) \kappa \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = \frac{(\omega+1)}{k\left(\frac{\omega}{k}+1\right)\epsilon}$
 732 will be necessary.

733 C.1.3 Main result for the general convex setting

734 **Theorem 6** (Convex). For a general convex function $f(\mathbf{w})$ with optimal solution $\mathbf{w}^{(*)}$, using
 735 *FedSKETCH*(τ, η, γ) to optimize $\tilde{f}(\mathbf{w}, \phi) = f(\mathbf{w}) + \frac{\phi}{2} \|\mathbf{w}\|^2$, for all $0 \leq t \leq R\tau - 1$, under
 736 Assumptions 1 to 2, if the learning rate satisfies

$$1 \geq \tau^2 L^2 \eta^2 + \left(\frac{\omega}{k} + 1\right) \eta \gamma L \tau$$

737 and if the all the models initiate with $\mathbf{w}^{(0)}$, with $\phi = \frac{1}{\sqrt{k\tau}}$ and $\eta = \frac{1}{2L\gamma\tau\left(1+\frac{\omega}{k}\right)}$ we obtain:

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] &\leq e^{-\frac{R}{2L\left(1+\frac{\omega}{k}\right)\sqrt{m\tau}}} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) \\ &\quad + \left[\frac{\sqrt{k}\sigma^2}{8\sqrt{\tau}\gamma^2\left(1+\frac{\omega}{k}\right)^2} + \frac{(\omega+1)\sigma^2}{4\left(\frac{\omega}{k}+1\right)\sqrt{k\tau}}\right] + \frac{1}{2\sqrt{k\tau}} \|\mathbf{w}^{(*)}\|^2 \end{aligned} \quad (26)$$

738 We note that above theorem implies that to achieve a convergence error of ϵ we need to have
 739 $R = O\left(L\left(1+\frac{\omega}{k}\right)\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{(\omega+1)^2}{k\left(\frac{\omega}{k}+1\right)^2\epsilon}\right)$.

740 *Proof.* Since $\tilde{f}(\mathbf{w}^{(r)}, \phi) = f(\mathbf{w}^{(r)}) + \frac{\phi}{2} \|\mathbf{w}^{(r)}\|^2$ is ϕ -PL, according to Theorem 5, we have:

$$\begin{aligned} &\tilde{f}(\mathbf{w}^{(R)}, \phi) - \tilde{f}(\mathbf{w}^{(*)}, \phi) \\ &= f(\mathbf{w}^{(r)}) + \frac{\phi}{2} \|\mathbf{w}^{(r)}\|^2 - \left(f(\mathbf{w}^{(*)}) + \frac{\phi}{2} \|\mathbf{w}^{(*)}\|^2\right) \\ &\leq (1 - \eta\gamma\phi\tau)^R \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\phi} \left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + (1+\omega)\frac{\gamma\eta L\sigma^2}{2k}\right] \end{aligned} \quad (27)$$

741 Next rearranging (27) and replacing μ with ϕ leads to the following error bound:

$$\begin{aligned} &f(\mathbf{w}^{(R)}) - f^* \\ &\leq (1 - \eta\gamma\phi\tau)^R \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\phi} \left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + (1+\omega)\frac{\gamma\eta L\sigma^2}{2k}\right] \\ &\quad + \frac{\phi}{2} \left(\|\mathbf{w}^*\|^2 - \|\mathbf{w}^{(r)}\|^2\right) \\ &\leq e^{-(\eta\gamma\phi\tau)R} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \frac{1}{\phi} \left[\frac{1}{2}L^2\tau\eta^2\sigma^2 + (1+\omega)\frac{\gamma\eta L\sigma^2}{2k}\right] + \frac{\phi}{2} \|\mathbf{w}^{(*)}\|^2 \end{aligned}$$

742 Next, if we set $\phi = \frac{1}{\sqrt{k\tau}}$ and $\eta = \frac{1}{2\left(1+\frac{\omega}{k}\right)L\gamma\tau}$, we obtain that

$$\begin{aligned} &f(\mathbf{w}^{(R)}) - f^* \\ &\leq e^{-\frac{R}{2\left(1+\frac{\omega}{k}\right)L\sqrt{m\tau}}} \left(f(\mathbf{w}^{(0)}) - f(\mathbf{w}^{(*)})\right) + \sqrt{k\tau} \left[\frac{\sigma^2}{8\tau\gamma^2\left(1+\frac{\omega}{k}\right)^2} + \frac{(\omega+1)\sigma^2}{4\left(\frac{\omega}{k}+1\right)\tau k}\right] + \frac{1}{2\sqrt{k\tau}} \|\mathbf{w}^{(*)}\|^2, \end{aligned}$$

743 thus the proof is complete. \square

C.2 Proof of Theorem 2

The proof of Theorem 2 follows directly from the results in [13]. We first mention the general Theorem 7 from [13] for general compression noise ω . Next, since the sketching PRIVIX and HEAPRIX, satisfy Assumption 4 with $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively with probability $1 - \frac{\delta}{R}$ per communication round, all the results in Theorem 2, conclude from Theorem 7 with probability $1 - \delta$ (by taking union over the all probabilities of each communication rounds with probability $1 - \delta/R$) and plugging $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively into the corresponding convergence bounds. For the heterogeneous setting, the results in [13] requires the following extra assumption that naturally holds for the sketching:

Assumption 5 ([13]). *The compression scheme Q for the heterogeneous data distribution setting satisfies the following condition $\mathbb{E}_Q[\|\frac{1}{m} \sum_{j=1}^m Q(\mathbf{x}_j)\|^2 - \|Q(\frac{1}{m} \sum_{j=1}^m \mathbf{x}_j)\|^2] \leq G_q$.*

We note that since sketching is a linear compressor, in the case of our algorithms for heterogeneous setting we have $G_q = 0$.

Next, we restate the Theorem in [13] here as follows:

Theorem 7. *Consider FedCOMGATE in [13]. If Assumptions 1, 3, 4 and 5 hold, then even for the case the local data distribution of users are different (heterogeneous setting) we have*

- **non-convex:** By choosing stepsizes as $\eta = \frac{1}{L\gamma} \sqrt{\frac{p}{R\tau(\omega+1)}}$ and $\gamma \geq p$, we obtain that the iterates satisfy $\frac{1}{R} \sum_{r=0}^{R-1} \|\nabla f(\mathbf{w}^{(r)})\|_2^2 \leq \epsilon$ if we set $R = O\left(\frac{\omega+1}{\epsilon}\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$.
- **Strongly convex or PL:** By choosing stepsizes as $\eta = \frac{1}{2L(\frac{\omega}{p}+1)\tau\gamma}$ and $\gamma \geq \sqrt{p\tau}$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left((\omega+1)\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$.
- **Convex:** By choosing stepsizes as $\eta = \frac{1}{2L(\omega+1)\tau\gamma}$ and $\gamma \geq \sqrt{p\tau}$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left(\frac{L(1+\omega)}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{p\epsilon^2}\right)$.

Proof. Since the sketching methods PRIVIX and HEAPRIX, satisfy the Assumption 4 with $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively with probability $1 - \frac{\delta}{R}$ per communication round, we conclude the proofs of Theorem 2 using Theorem 7 with probability $1 - \delta$ (by taking union over all communication rounds) and plugging $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively into the convergence bounds. \square

771 **D Numerical Experiments and Additional Results**

772 **D.1 Implementation of FetchSGD**

773 Our implementation of FetchSGD basically follows the original paper (Algorithm 1 in [37]). The
774 only difference is that, in the original algorithm, the local workers compress the gradient (in every
775 local step) and transmit it to the central server. In our setting, we extend to the case with multiple local
776 updates, where the difference in local weights are transmitted (same as the standard FL framework).
777 Also, TopK compression is used to decode the sketches at the central server. We apply the same
778 implementation trick that when accumulating the errors, we only count the non-zero coordinates and
779 leave other coordinates zero for the accumulator. This greatly improves the empirical performance.

780 **D.2 Additional Plots for the MNIST Experiments**

781 **D.2.1 Homogeneous setting**

782 In the homogeneous case, each node has same data distribution. To achieve this setting, we randomly
783 choose samples uniformly from 10 classes of hand-written digits. The train loss and test accuracy
784 are provided in Figure 3, where we report local epochs $\tau = 2$ in addition to the main context (single
785 local update). The number of users is set to 50, and in each round of training we randomly pick half
786 of the nodes to be active (i.e., receiving data and performing local updates). We can draw similar
787 conclusion: FS-HEAPRIX consistently performs better than other competing methods. The test
788 accuracy increases with larger τ in homogeneous setting.

789 **D.2.2 Heterogeneous setting**

790 Analogously, we present experiments on MNIST dataset under heterogeneous data distribution,
791 including $\tau = 2$. We simulate the setting by only sending samples from one digit to each local
792 worker (very few nodes get two classes). We see from Figure 4 that FS-HEAPRIX shows consistent
793 advantage over competing methods. SketchedSGD performs poorly in this case.

794 **D.3 Additional Experiments: CIFAR-10**

795 We conduct similar sets of experiments on CIFAR10 dataset. We also use the simple LeNet CNN
796 structure, as in practice small models are more favorable in federated learning, due to the limitation of
797 mobile devices. The test accuracy is presented in Figure 5 and Figure 6, for respectively homogeneous
798 and heterogeneous data distribution. In general, we retrieve similar information as from MNIST
799 experiments: our proposed FS-HEAPRIX improves FS-PRIVIX and SketchedSGD in all cases. We
800 note that although the test accuracy provided by LeNet cannot reach the state-of-the-art accuracy
801 given by some huge models, it is also informative in terms of comparing the relative performance of
802 different sketching methods.

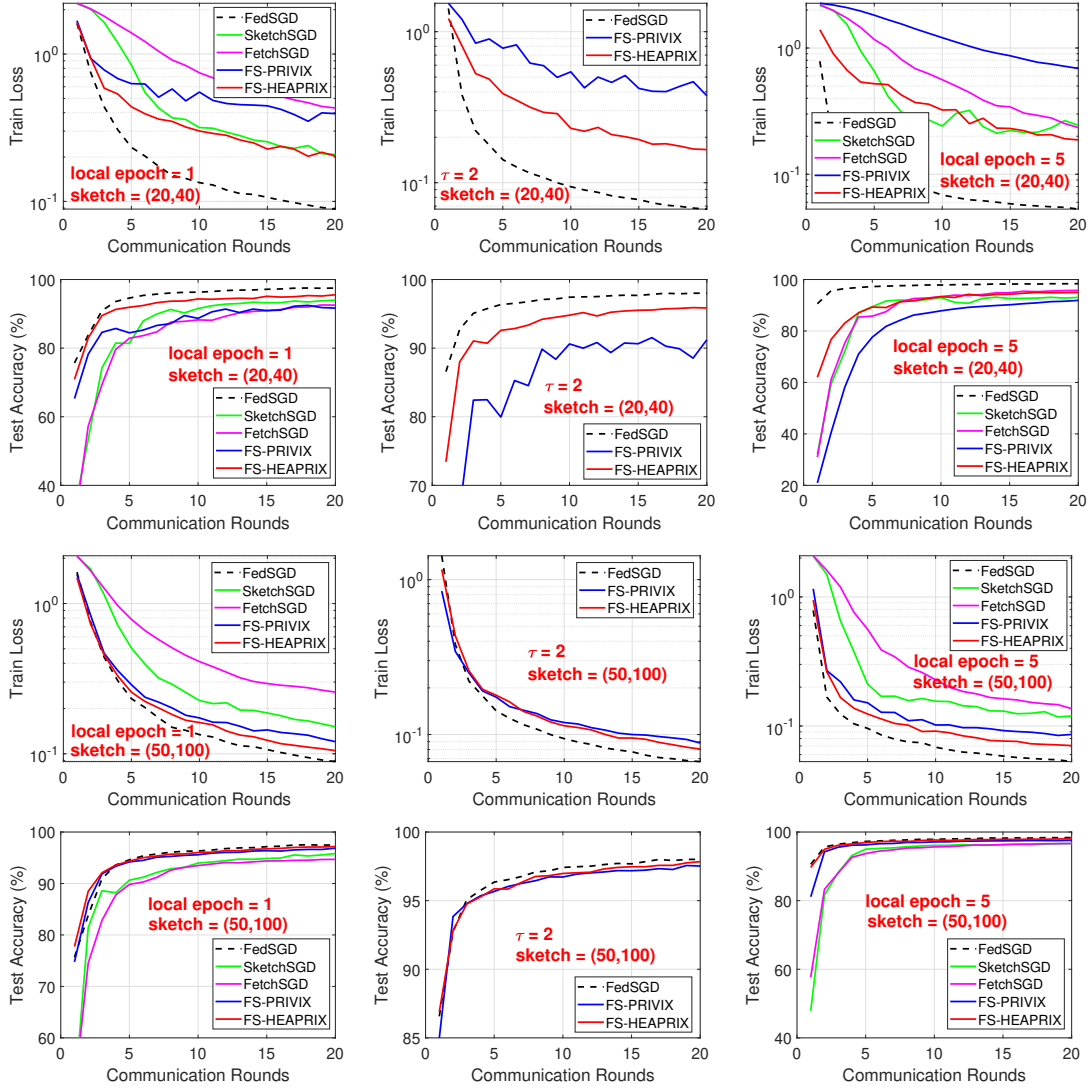


Figure 3: MNIST Homogeneous case: Comparison of compressed optimization methods on LeNet CNN architecture.

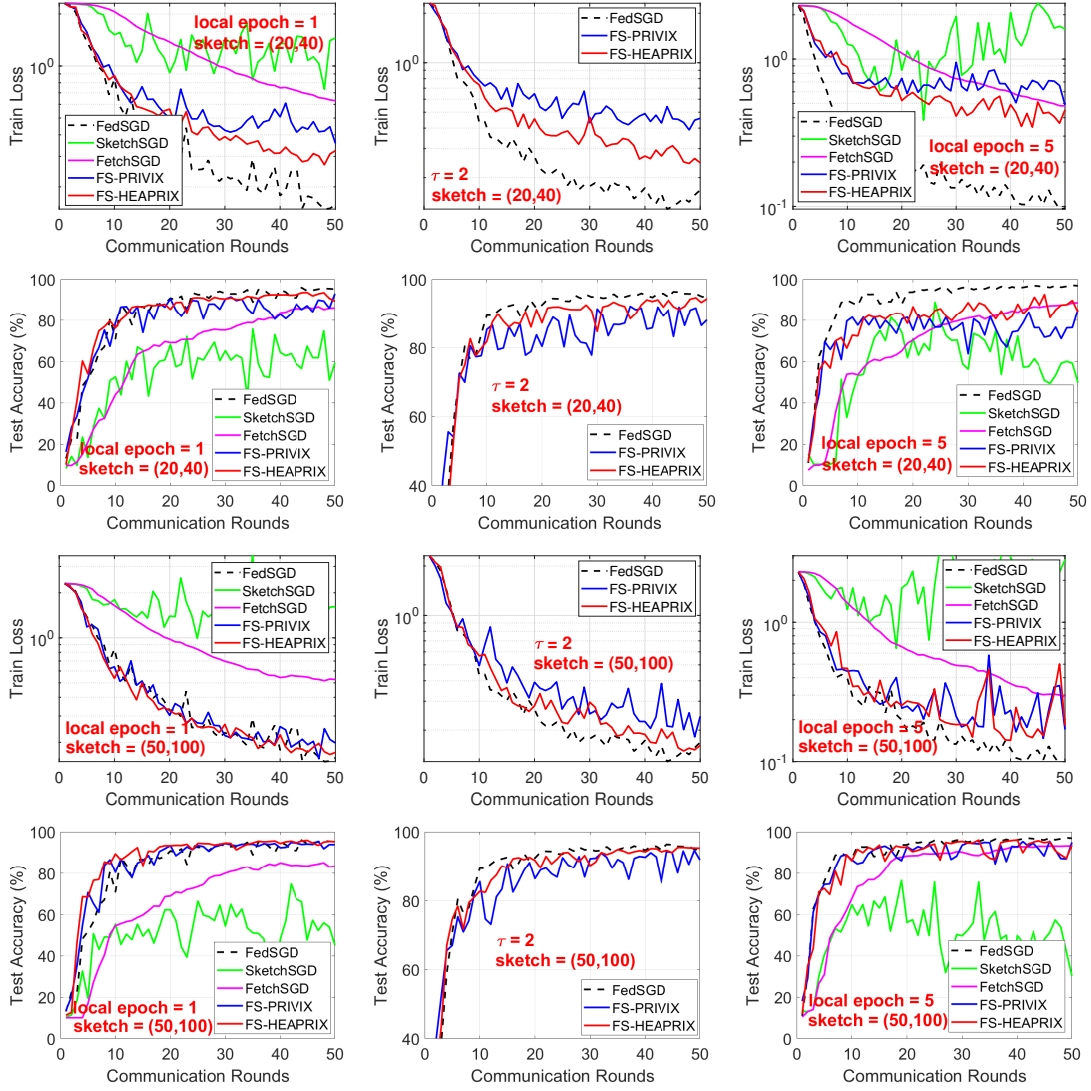


Figure 4: MNIST Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN architecture.

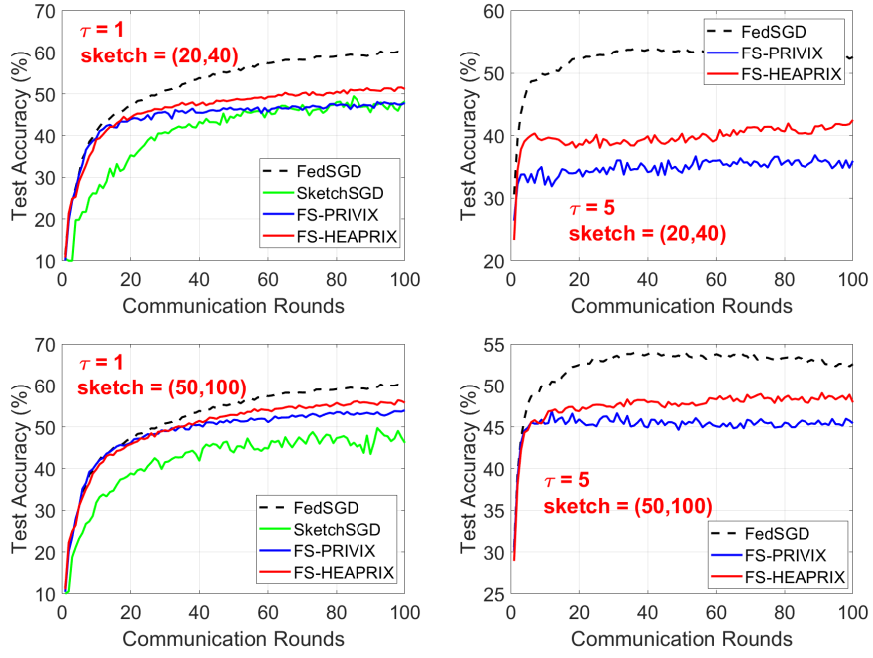


Figure 5: Homogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.

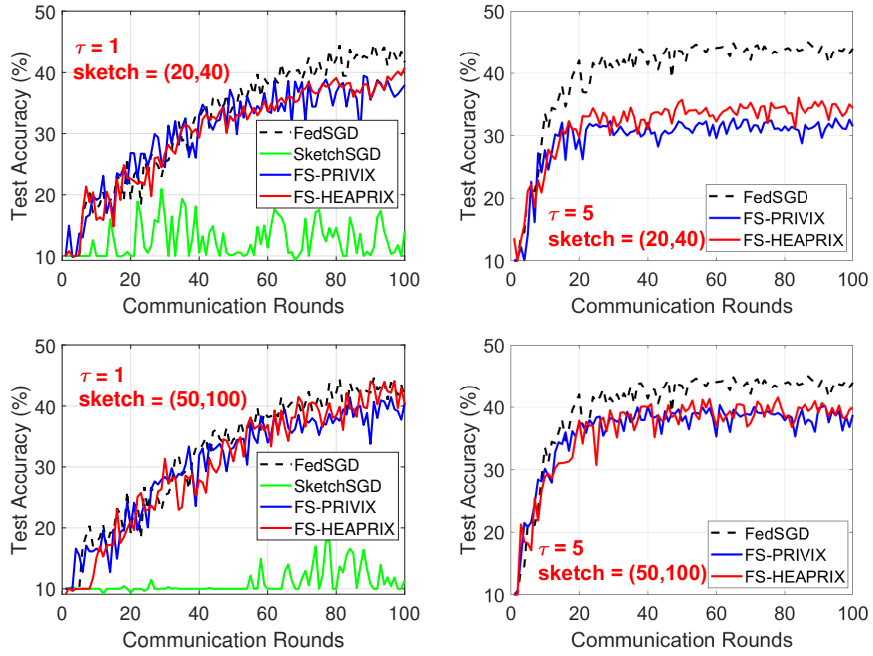


Figure 6: Heterogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.