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 C of this supplementary file, a through comparison with notable related works. Section D contains the proofs of our results and Section E presents additional numerical runs.

A Notations and Definitions

Notation. Here we denote the count sketch of the vector $\boldsymbol{x}$ by $\mathbf{S}(\boldsymbol{x})$ and with an abuse of notation, we indicate the expectation over the randomness of count sketch with $\mathbb{E}_{\mathbf{S}}[.]$. 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 $\mathbb{E}_{\mathcal{K}}[.]$.

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
$\mathbf{S}(\boldsymbol{x})$	$\triangleq$	Count sketch of the vector $\boldsymbol{x}$
$\mathbb{U}(.)$	$\triangleq$	Class of unbiased compressor, see Definition 1

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

A.1 Count sketch

In this paper, we exploit the commonly used `Count Sketch` [Charikar et al., 2004] which uses two sets of functions that encode any input vector $\boldsymbol{x}$ **into a hash table** $\boldsymbol{S}_{m\times t}(\boldsymbol{x})$. Pairwise independent hash functions $\{h_{j,1\leq j\leq t} : [d] \to m\}$ are used along with another set of pairwise independent sign hash functions $\{\text{sign}_{j,1\leq j\leq t} : [d] \to \{+1, -1\}\}$ to map entries of $\boldsymbol{x}$ (x_i, $1 \leq i \leq d$) into t different columns of $\mathbf{S}_{m\times t}$, wherein, to lower the dimension of the input vector, we usually have $d \gg mt$. The final update reads $\mathbf{S}[j][h_j(i)] = \mathbf{S}[j][h_j(i)] + \text{sign}_j(i)x_i$ for any $1 \leq j \leq t$. Generating compressed output is described in Algorithm 5.

Algorithm 5 Count Sketch (CS) [Charikar et al., 2004]

Inputs: $\boldsymbol{x} \in \mathbb{R}^d, t, k, \mathbf{S}_{m\times t}, h_j(1 \le i \le t), \text{sign}_j(1 \le i \le t)$
Compress vector $\boldsymbol{x} \in \mathbb{R}^d$ **into** $\mathbf{S}(\boldsymbol{x})$**:**
for $\boldsymbol{x}_i \in \boldsymbol{x}$ **do**
  for $j = 1, \cdots, t$ **do**
    $\mathbf{S}[j][h_j(i)] = \mathbf{S}[j-1][h_{j-1}(i)] + \text{sign}_j(i).\boldsymbol{x}_i$
  end for
end for
return $\mathbf{S}_{m\times t}(\boldsymbol{x})$

A.2 PRIVIX method and compression error of HEAPRIX

For the sake of completeness we review PRIVIX algorithm that is also mentioned in [Li et al., 2019] as follows:

Algorithm 6 PRIVIX/DiffSketch [Li et al., 2019]: Unbiased compressor based on sketching.

Inputs: $\boldsymbol{x} \in \mathbb{R}^d, t, m, \mathbf{S}_{m\times t}, h_j(1 \le i \le t), sign_j(1 \le i \le t)$
Query $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ **from** $\mathbf{S}(\mathbf{x})$**:**
for $i = 1, \ldots, d$ **do**
  $\tilde{\boldsymbol{x}}[i] = \text{Median}\{\text{sign}_j(i).\mathbf{S}[j][h_j(i)] : 1 \le j \le t\}$
end for
Output: $\tilde{\boldsymbol{x}}$

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

Corollary 2. *Based on Theorem 3 of [Horváth and Richtárik, 2020] and using Algorithm 2, we have* $\mathcal{C}(x) \in \mathbb{U}(c\frac{d}{m})$. *This shows that unlike PRIVIX (Algorithm 6) the compression noise can be made as small as possible using large size of hash table.*

Proof. The proof simply follows from Theorem 3 in [Horváth and Richtárik, 2020] and Algorithm 2 by setting $\Delta_1 = c\frac{d}{m}$ and $\Delta_2 = 1 + c\frac{d}{m}$ we obtain $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} = c\frac{d}{m} = O\left(\frac{d}{m}\right)$ for the compression error of HEAPRIX. □

B Convergence of FedSketchGate for Convex Objectives

Theorem 3. *Suppose Assumptions 1-2 hold. Given* $0 < m \leq d$ *and considering Algorithm 3 with sketch size* $B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ *and* $\gamma \geq k$*, with probability* $1-\delta$ *for* ***Convex*** *case,* $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ *satisfies* $\mathbb{E}\Big[f(\boldsymbol{x}^{(R-1)}) - f(\boldsymbol{x}^{(*)})\Big] \leq \epsilon$ *if we set:*

- *FS-PRIVIX, for* $\eta = \frac{1}{2L(cd/m+1/k)\tau\gamma}$*:* $R = O\left(L\left(1/k + d/m\right)/\epsilon \log\left(1/\epsilon\right)\right)$ *and* $\tau = O\left(1/\epsilon^2\right)$.
- *FS-HEAPRIX, for* $\eta = \frac{1}{2L((cd-m)/mk+1)\tau\gamma}$*:* $R = O\left(L\left(1/k + (d-m)/m\right)/\epsilon \log\left(1/\epsilon\right)\right)$ *and* $\tau = O\left(1/\epsilon^2\right)$.

Theorem 4. *Suppose Assumptions 1 and 3 hold. Given* $0 < m \leq d$*, and considering* FedSKETCHGATE *in Algorithm 4 with sketch size* $B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ *and* $\gamma \geq p$ *with probability* $1-\delta$ *for* ***convex*** *case,* $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ *satisfies* $\mathbb{E}[f(\boldsymbol{x}^{(R-1)}) - f(\boldsymbol{x}^{(*)})] \leq \epsilon$ *if:*

- *FS-PRIVIX, for* $\eta = 1/(2L(cd/m+1)\tau\gamma)$*:* $R = O\left(L(d/m+1)\epsilon \log(1/\epsilon)\right)$ *and* $\tau = O\left(1/(p\epsilon^2)\right)$.
- *FS-HEAPRIX, for* $\eta = m/(2Lcd\tau\gamma)$*:* $R = O\left(L(d/m)\epsilon \log(1/\epsilon)\right)$ *and* $\tau = O\left(1/(p\epsilon^2)\right)$.

C Summary of comparison of our results with prior works

For the purpose of further clarification, we summarize the comparison of our results with related works. We recall that p is the number of devices, d is the dimension of the model, κ is the condition number, ϵ is the target accuracy, R is the number of communication rounds, and τ is the number of local updates. We start with the homogeneous setting comparison. Comparison of our results and 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. Here, p is the number of devices, μ is the PL constant, m is the number of bins of hash tables, d is the dimension of the model, κ is the condition number, ϵ is the target accuracy, R is the number of communication rounds, and τ is the number of local updates. UG and PP stand for Unbounded Gradient and Privacy Property respectively.

Reference	Non-Convex	UG	PP
[Li et al., 2019]	–	–	$R = O\left(\frac{\mu^2 d}{\epsilon^2}\right),\ \tau = 1$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pRB = O\left(\frac{p\mu^2 d}{\epsilon^2} k \log\left(\frac{\mu^2 d^2}{\epsilon^2 \delta}\right)\right)$
Ivkin et al. [Ivkin et al., 2019]	$R = O\left(\max\left(\frac{d}{m\sqrt{\epsilon}}, \frac{1}{\epsilon}\right)\right),\ \tau = 1,\ B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ $pRB = O\left(\frac{pd}{m\epsilon} \log\left(\frac{d}{\delta\sqrt{\epsilon}} \max\left(\frac{d}{m}, \frac{1}{\sqrt{\epsilon}}\right)\right)\right)$	✗	✗
Theorem 1	$\boldsymbol{R = O\left(\frac{1}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left((\mu^2(cd-m) + \frac{\mu^2}{k})\frac{1}{\epsilon}\right)}$ $\boldsymbol{B = O(m \log\left(\frac{dR}{\delta}\right))}$ $\boldsymbol{kBR = O(mk/\epsilon \log\left(\frac{d}{\epsilon\delta}\right))}$	✔	✗

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. [Basu et al., 2019] (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. [Li et al., 2019]	–	$R{=}O\left(\frac{d}{m\epsilon^2}\right)$ $\tau{=}1$ $B = O\left(m\log\left(\frac{d^2}{m\epsilon^2\delta}\right)\right)$	✗	✔
Rothchild et al. [Rothchild et al., 2020]	$R = O\left(\max(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon})\right)$ $\tau = 1$ $B = O\left(m\log\left(\frac{d}{\delta}\max(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon})\right)\right)$ $RB = O\left(m\max(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon})\log\left(\frac{d}{\delta}\max(\frac{1}{\epsilon^2}, \frac{d^2-md}{m^2\epsilon})\right)\right)$	–	✗	✗
Rothchild et al. [Rothchild et al., 2020]	$R = O\left(\frac{\max(I^{2/3}, 2-\alpha)}{\epsilon^3}\right)$ $\tau = 1$ $B = O\left(\frac{m}{\alpha}\log\left(\frac{d\max(I^{2/3}, 2-\alpha)}{\epsilon^3\delta}\right)\right)$ $RB = O\left(\frac{m\max(I^{2/3}, 2-\alpha)}{\epsilon^3\alpha}\log\left(\frac{d\max(I^{2/3}, 2-\alpha)}{\epsilon^3\delta}\right)\right)$	–	✗	✗
Theorem 2	$\boldsymbol{R = O\left(\frac{d}{m\epsilon}\right)}$ $\boldsymbol{\tau = O(\frac{1}{p\epsilon})}$ $\boldsymbol{B = O(m\log(\frac{d^2}{m\epsilon\delta}))}$ $\boldsymbol{RB = O(\frac{d}{\epsilon}\log(\frac{d^2}{m\epsilon\delta}\log(\frac{1}{\epsilon})))}$	$\boldsymbol{R{=}O(\frac{d}{m\epsilon}\log(\frac{1}{\epsilon}))}$ $\boldsymbol{\tau{=}O(\frac{1}{p\epsilon^2})}$ $\boldsymbol{B = O(m\log(\frac{d^2}{m\epsilon\delta}))}$	✔	✔

Comparison with [Haddadpour et al., 2020] and [Reisizadeh et al., 2020] Convergence analysis of algorithms in [Haddadpour et al., 2020] relies on unbiased compression, while in this paper our FL algorithm based on `HEAPRIX` enjoys from unbiased compression with equivalent biased compression variance. Moreover, we highlight that the convergence analysis of `FedCOMGATE` is based on the extra assumption of boundedness of the difference between the average of compressed vectors and compressed averages of vectors. However, we do not need this extra assumption as it is satisfied naturally due to linearity of sketching. Finally, as pointed out in Remark 2, our algorithms enjoy from a bidirectional compression property, unlike `FedCOMGATE` in general. Furthermore, since results in [Haddadpour et al., 2020] improve the communication complexity of FedPAQ algorithm, developed in [Reisizadeh et al., 2020], hence `FedSKETCH` and `FedSKETCHGATE` improves the communication complexity obtained in [Reisizadeh et al., 2020].

[Basu et al., 2019]. We note that the algorithm in [Basu et al., 2019] uses a composed compression and quantization while our algorithm is solely based on compression. So, in order to compare with algorithms in [Basu et al., 2019] we only consider Qsparse-local-SGD with compression and we let compression factor $\gamma = \frac{m}{d}$ (to compare with the same compression ratio induced with sketch size of mt). For strongly convex objective in Qsparse-local-SGD to achieve convergence error of ϵ they require $R = O\left(\kappa\frac{d}{m\sqrt{\epsilon}}\right)$ and $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$, 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. Similarly, for non-convex objective [Basu et al., 2019] requires $R = O\left(\frac{d}{m\epsilon^{1.5}}\right)$ and $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$, which is 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 cost of increasing number of local updates (which scales down with number of devices, p). Additionally, we highlight that our `FedSKETCHGATE` exploits the gradient tracking idea to deal with data heterogeneity, while algorithms in [Basu et al., 2019] does not develop such mechanism and may suffer from poor convergence in heterogeneous setting. We also note that setting $\tau = 1$ and using top_m compressor, the QSPARSE-local-SGD algorithm becomes similar to distributed SGD with sketching as they both use the error feedback framework to improve the compression variance. Finally, since the average of sparse vectors may not be sparse in general the number of transmitted bits from server to devices in QSPARSE-Local-SGD in [Basu et al., 2019] may not be sparse in general ($B = O(d)$), however our algorithms enjoy from bidirectional compression properly due to lower dimension and linearity properties of sketching ($B = O(m\log(\frac{Rd}{\delta}))$). Therefore, the total number of bits per device for strongly convex and

non-convex objective is improved respectively from $RB = O\left(\kappa \frac{d^2}{m\sqrt{\epsilon}}\right)$ and $RB = O\left(\frac{d^2}{m\epsilon^{1.5}}\right)$ in [Basu et al., 2019] 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 $RB = O\left(\log(\frac{d^2}{m\epsilon\delta})\frac{d}{\epsilon}\right)$.

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

[Philippenko and Dieuleveut, 2020]. The reference [Philippenko and Dieuleveut, 2020] considers two-way compression from parameter server to devices and vice versa. They provide the convergence rate of $R = O\left(\frac{\omega^{\mathrm{Up}}\omega^{\mathrm{Down}}}{\epsilon^2}\right)$ for strongly-objective functions where ω^{Up} and ω^{Down} are uplink and downlink's compression noise (specializing to our case for the sake of comparison $\omega^{\mathrm{Up}} = \omega^{\mathrm{Down}} = \theta(d)$) for general heterogeneous data distribution. In contrast, while our algorithms are using bidirectional compression due to use of sketching for communication, our convergence rate for strongly-convex objective is $R = O(\kappa\mu^2 d \log\left(\frac{1}{\epsilon}\right))$ with probability $1 - \delta$.

We would like to also mention that there prior studies such as [Tang et al., 2019] and [Zheng et al., 2019] that analyze the two-way compression, but since [Philippenko and Dieuleveut, 2020] is the state-of-the-art on this topic we only compared our results with these papers.

D Theoretical Proofs

We will use the following fact (which is also used in [Li et al., 2020c, Haddadpour and Mahdavi, 2019]) in proving results.

Fact 5 ([Li et al., 2020c, Haddadpour and Mahdavi, 2019]). *Let $\{x_i\}_{i=1}^p$ denote any fixed deterministic sequence. We sample a multiset $\mathcal{P}$ (with size K) uniformly at random where x_j is sampled with probability q_j for $1 \leq j \leq p$ with replacement. Let $\mathcal{P} = \{i_1, \ldots, i_K\} \subset [p]$ (some i_js may have the same value). Then*

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

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

Assumption 4 ([Haddadpour et al., 2020]). *The output of the compression operator $Q(\boldsymbol{x})$ is an unbiased estimator of its input $\boldsymbol{x}$, and its variance grows with the squared of the squared of ℓ_2-norm of its argument, i.e., $\mathbb{E}\left[Q(\boldsymbol{x})\right] = \boldsymbol{x}$ and $\mathbb{E}\left[\left\|Q(\boldsymbol{x}) - \boldsymbol{x}\right\|^2\right] \leq \omega \left\|\boldsymbol{x}\right\|^2$.*

We note that 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. Therefore, all the results in Theorem 1, by taking union over the all probabilities of each communication rounds, are concluded with probability $1 - \delta$ by plugging $\omega = c\frac{d}{m}$ and $\omega = c\frac{d}{m} - 1$ respectively into the corresponding convergence bounds.

D.1 Proof of Theorem 1

In this section, we study the convergence properties of our `FedSKETCH` method presented in Algorithm 3. Before developing the proofs for `FedSKETCH` in the homogeneous setting, we first mention the following intermediate lemmas.

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

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

Proof.

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

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

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

Next we show that from Assumptions 3, we have

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

To do so, note that

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

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

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

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

Further note that we have

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

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

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

and the proof is complete. □

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

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

Proof. We have:

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

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. □

The following lemma bounds the distance of local solutions from global solution at rth communication round.

Lemma 3. *Under Assumptions 2 we have:*

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

Proof. Note that

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\boldsymbol{w}_j^{(c,r)}\right\|_2^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{w}^{(r)}-\left(\boldsymbol{w}^{(r)}-\eta\sum_{k=0}^c\tilde{\mathbf{g}}_j^{(k,r)}\right)\right\|_2^2\right]\\
&= \mathbb{E}\left[\left\|\eta\sum_{k=0}^c\tilde{\mathbf{g}}_j^{(k,r)}\right\|_2^2\right]
\end{aligned}$$

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

where ① comes from $\mathbb{E}\left[\mathbf{x}^2\right]=\mathrm{Var}\left[\mathbf{x}\right]+\left[\mathbb{E}\left[\mathbf{x}\right]\right]^2$ and ② holds because $\mathrm{Var}\left(\sum_{j=1}^{n}\mathbf{x}_j\right)=\sum_{j=1}^{n}\mathrm{Var}\left(\mathbf{x}_j\right)$ for i.i.d. vectors $\mathbf{x}_i$ (and i.i.d. assumption comes from i.i.d. sampling), and finally ③ follows from Assumption 2. □

D.1.1 Main result for the non-convex setting

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

Theorem 6 (non-convex)**.** *For* `FedSKETCH`(τ,η,γ)*, for all* $0\leq t\leq R\tau-1$*, under Assumptions 1 to 2, if the learning rate satisfies*

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

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

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

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

Proof. Before proceeding with the proof of Theorem 6, we would like to highlight that

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

From the updating rule of Algorithm 3 we have

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

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

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

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

Then using the unbiased estimation property of sketching we have:

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

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

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

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

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

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

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

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

$$1 \geq \tau L^2\eta^2\tau + (\omega+1)\eta\gamma L\tau .$$

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

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

From the above inequality, is it easy to see that in order to achieve a linear speed up, we need to have $\eta\gamma = O\left(\frac{1}{\sqrt{R\tau(\omega+\frac{1}{k})}}\right)$. □

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

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

Note that according to (19), if we pick a fixed constant value for γ*, in order to achieve an* ϵ*-accurate solution,* $R = O\left(\frac{1}{\epsilon}\right)$ *communication rounds and* $\tau = O\left(\frac{\omega+\frac{1}{k}}{\epsilon}\right)$ *local updates are necessary.*

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

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

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

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

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

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

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

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

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

which matches the rate obtained in [Wang and Joshi, 2018]. In this case the communication complexity and the number of local updates become

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

which simply implies that in this special case the convergence rate of our algorithm reduces to the rate obtained in [Wang and Joshi, 2018], which indicates the tightness of our analysis.

D.1.2 Main result for the PL/Strongly convex setting

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

Theorem 7 (PL or strongly convex)**.** *For* `FedSKETCH`(τ, η, γ)*, for all* $0 \leq t \leq R\tau - 1$*, under Assumptions 1 to 2 and 3, if the learning rate satisfies*

$$1 \geq \tau^2 L^2 \eta^2 + (\omega + 1)\, \eta\gamma L\tau$$

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

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

Proof. From (18) under condition:

$$1 \geq \tau L^2\eta^2\tau + (\omega + 1)\eta\gamma L\tau$$

we obtain:

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

which leads to the following bound:

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

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

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

□

Corollary 5. *If we let* $\eta\gamma\mu\tau \leq \frac{1}{2}$*,* $\eta = \frac{1}{2L\left(\omega + \frac{1}{k}\right)\tau\gamma}$ *and* $\kappa = \frac{L}{\mu}$ *the convergence error in Theorem 7, with* $\gamma \geq k$ *results in:*

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

which indicates that to achieve an error of ϵ*, we need to have* $R = O\left(\left(\omega + \frac{1}{k}\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = \frac{(\omega+1)}{\left(\omega + \frac{1}{k}\right)\epsilon}$.

D.1.3 Main result for the general convex setting

Theorem 8 (Convex)**.** *For a general convex function $f(\boldsymbol{w})$ with optimal solution $\boldsymbol{w}^{(*)}$, using* `FedSKETCH`(τ, η, γ) *to optimize $\tilde{f}(\boldsymbol{w}, \phi) = f(\mathbf{w}) + \frac{\phi}{2}\left\|\boldsymbol{w}\right\|^2$, for all $0 \leq t \leq R\tau - 1$, under Assumptions 1 to 2, if the learning rate satisfies*

$$1 \geq \tau^2 L^2 \eta^2 + (\omega + 1)\,\eta\gamma L\tau$$

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

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

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

Proof. Since $\tilde{f}(\boldsymbol{w}^{(r)}, \phi) = f(\boldsymbol{w}^{(r)}) + \frac{\phi}{2}\left\|\boldsymbol{w}^{(r)}\right\|^2$ is ϕ-PL, according to Theorem 7, we have:

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

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

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

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

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

thus the proof is complete. □

D.2 Proof of Theorem 2

The proof of Theorem 2 follows directly from the results in [Haddadpour et al., 2020]. We first mention the general Theorem 9 from [Haddadpour et al., 2020] 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 9 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 [Haddadpour et al., 2020] requires the following extra assumption that naturally holds for the sketching:

Assumption 5 ([Haddadpour et al., 2020])**.** *The compression scheme Q for the heterogeneous data distribution setting satisfies the following condition $\mathbb{E}_Q[\|\frac{1}{m}\sum_{j=1}^m Q(\boldsymbol{x}_j)\|^2 - \|Q(\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_j)\|^2] \leq G_q$.*

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 [Haddadpour et al., 2020] here as follows:

Theorem 9. *Consider `FedCOMGATE` in [Haddadpour et al., 2020]. 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 $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ 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 $\eta = \frac{1}{2L\left(\frac{\omega}{p}+1\right)\tau\gamma}$ and $\gamma \geq \sqrt{p\tau}$, we obtain that the iterates satisfy $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq \epsilon$ 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 $\mathbb{E}\Big[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\Big] \leq \epsilon$ 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 9 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. □

E Numerical Experiments and Additional Results

E.1 Implementation of FetchSGD

Our implementation of `FetchSGD` basically follows the original paper (Algorithm 1 in [Rothchild et al., 2020]). The only difference is that, in the original algorithm, the local workers compress the gradient (in every local step) and transmit it to the central server. In our setting, we extend to the case with multiple local updates, where the difference in local weights are transmitted (same as the standard FL framework). Also, TopK compression is used to decode the sketches at the central server. We apply the same implementation trick that when accumulating the errors, we only count the non-zero coordinates and leave other coordinates zero for the accumulator. This greatly improves the empirical performance.

E.2 Additional Plots for the MNIST Experiments

E.2.1 Homogeneous setting

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

E.2.2 Heterogeneous setting

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

E.3 Additional Experiments: CIFAR-10

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

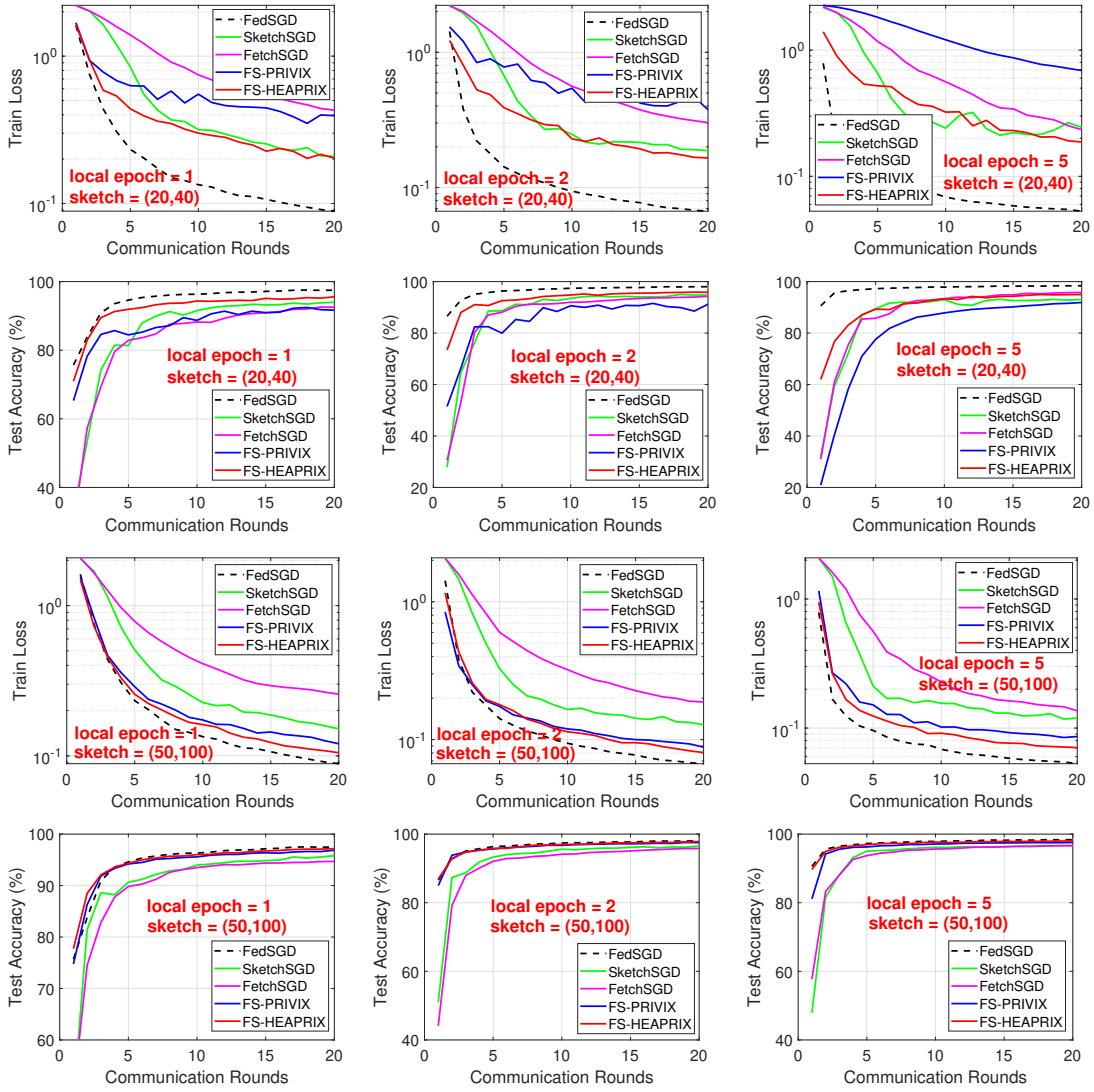

Figure 3: MNIST Homogeneous case: Comparison of compressed optimization methods on LeNet CNN architecture.

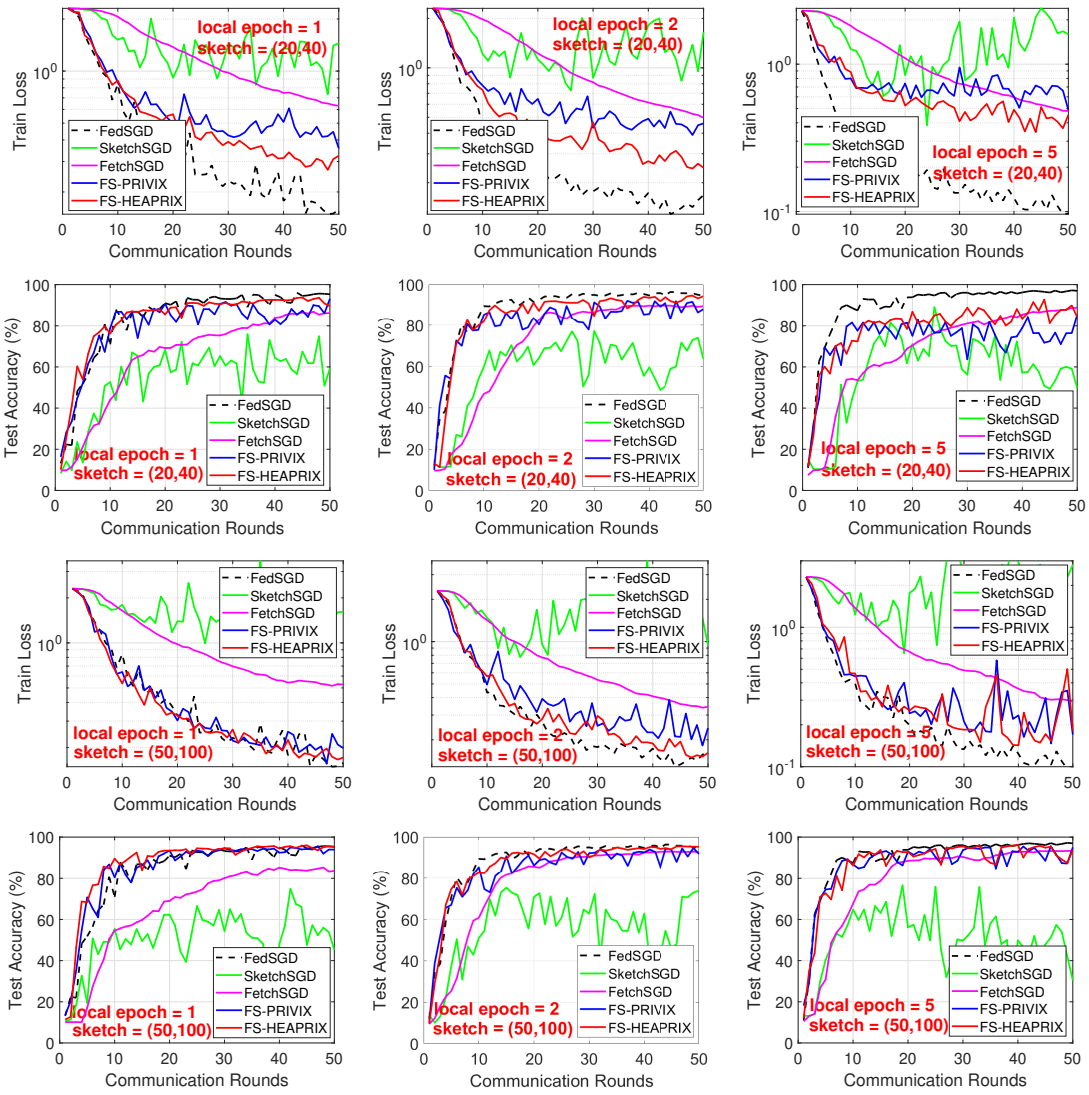

Figure 4: MNIST Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN architecture.

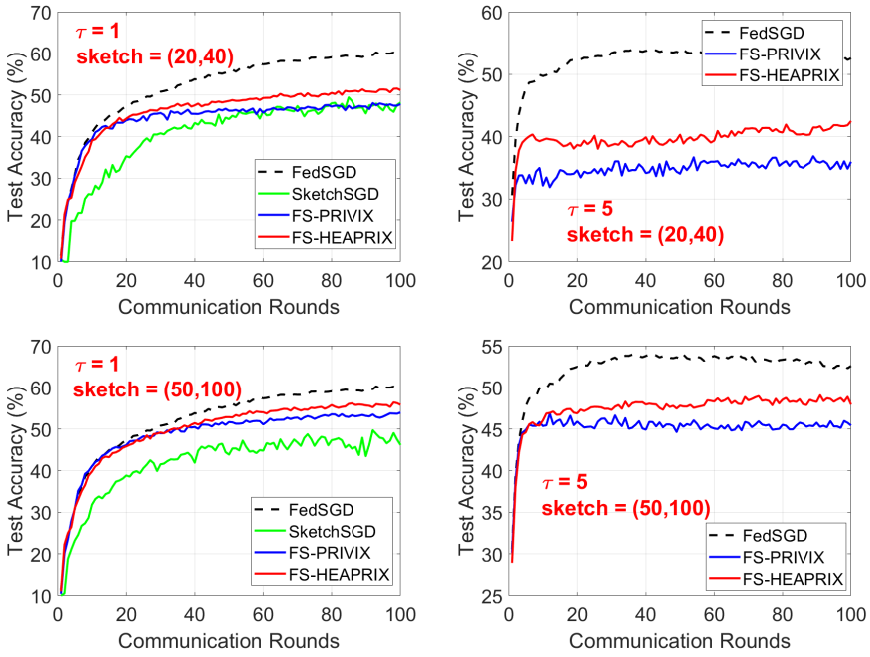

Figure 5: Homogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.

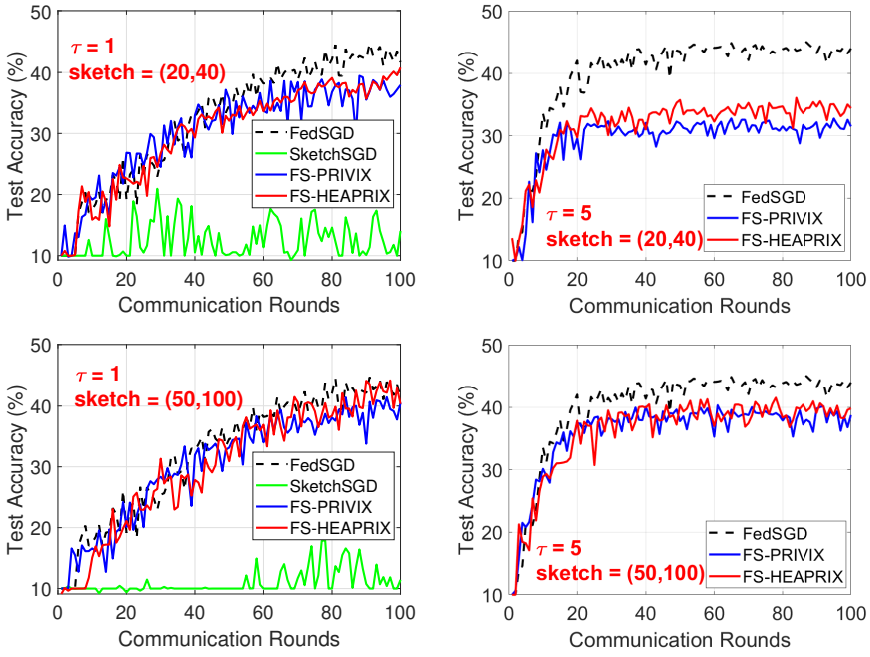

Figure 6: Heterogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.