Appendix for FedSKETCH: Communication-Efficient Federated Learning via Sketching

A Notations and Defintions

Notation. Here we indicate the count sketch of the vector $\boldsymbol{x}$ with $\mathbf{S}(\boldsymbol{x})$ and with 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 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
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 $\boldsymbol{x}$
$\mathbb{U}(\Delta)$	$\triangleq$	Class of unbiased compressor, see Definition 1

Definition 3. *A randomized mechanism $\mathcal{O}$ satisfies $\epsilon-$differential privacy, if for input data S_1 and S_2 differing by up to one point, and for output D of $\mathcal{O}$,*

$$\Pr\left[\mathcal{O}(S_1) \in D\right] \leq \exp\left(\epsilon\right) \Pr\left[\mathcal{O}(S_2) \in D\right] .$$

For smaller ϵ, it becomes difficult to specify the input data, hence, implying stronger privacy.

A.1 PRIVIX and compression error of HEAPRIX

For the sake of completeness we review PRIVIX algorithm that is also mentioned in [27] as follows:

Algorithm 6 PRIVIX [27]: Unbiased compressor based on sketching.

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

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

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. [19]	$R = O\left(\max\left(\frac{\mu^2 d}{\sqrt{\epsilon}}, \frac{1}{\epsilon}\right)\right),\ \tau = 1,\ B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ $pRB = O\left(\frac{p\mu^2 d}{\epsilon} m \log\left(\frac{d}{\delta\sqrt{\epsilon}} \max\left(\mu^2 d, \frac{1}{\sqrt{\epsilon}}\right)\right)\right)$	✗	✗
Theorem 1	$\boldsymbol{R = O\left(\kappa\left(\frac{\mu^2 d - 1}{k} + 1\right) \log\left(\frac{1}{\epsilon}\right)\right),\ \tau = O\left(\frac{(\mu^2 d)}{k\left(\frac{\mu^2 d}{k}+1\right)\epsilon}\right), B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{kBR = O\left(m\kappa(\mu^2 d - 1 + k) \log\frac{1}{\epsilon} \log\left(\frac{\kappa(d\frac{\mu^2 d - 1}{k} + d) \log\frac{1}{\epsilon}}{\delta}\right)\right)}$	✔	✔

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
Li et al. [27]	–	$R{=}O\left(\frac{\mu^2 d}{\epsilon^2}\right)$ $\tau{=}1$ $B = O\left(m \log\left(\frac{\mu^2 d^2}{\epsilon^2 \delta}\right)\right)$	✗	✔
Rothchild et al. [36]	$R = O\left(\max(\frac{1}{\epsilon^2}, \frac{d^2 - md}{m^2 \epsilon})\right)$ $\tau = 1$ $B = O\left(m \log\left(\frac{d}{\epsilon^2 \delta}\right)\right)$ $BR = O\left(\frac{m}{\epsilon^2} \max(\frac{1}{\epsilon^2}, \frac{d^2 - md}{m^2\epsilon}) \log\left(\frac{d}{\delta} \max(\frac{1}{\epsilon^2}, \frac{d^2 - md}{m^2 \epsilon})\right)\right)$	–	✗	✗
Rothchild et al. [36]	$R = O\left(\frac{\max(I^{2/3}, 2-\alpha)}{\epsilon^3}\right)$ $\tau = 1$ $B = O\left(\frac{m}{\alpha} \log\left(\frac{d \max(I^{2/3}, 2-\alpha)}{\epsilon^3 \delta}\right)\right)$ $BR = O\left(\frac{m \max(I^{2/3}, 2-\alpha)}{\epsilon^3 \alpha} \log\left(\frac{d \max(I^{2/3}, 2-\alpha)}{\epsilon^3 \delta}\right)\right)$	–	✗	✗
Theorem 2	$\boldsymbol{R = O\left(\frac{\mu^2 d}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left(\frac{1}{p\epsilon}\right)}$ $\boldsymbol{B = O\left(m \log\left(\frac{\mu^2 d^2}{\epsilon\delta}\right)\right)}$ $\boldsymbol{BR = O\left(\frac{m(\mu^2 d)}{\epsilon} \log\left(\frac{\mu^2 d^2}{\epsilon\delta} \log\left(\frac{1}{\epsilon}\right)\right)\right)}$	$\boldsymbol{R{=}O\left(\frac{\mu^2 d}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau{=}O\left(\frac{1}{p\epsilon^2}\right)}$ $\boldsymbol{B = O\left(m \log\left(\frac{\mu^2 d^2}{\epsilon\delta}\right)\right)}$	✔	✔

Corollary 2. *Based on [18, Theorem 3] and using Algorithm 3, we have $C(x) \in \mathbb{U}(\mu^2 d)$. 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 [18] and Algorithm 3 by setting $\Delta_1 = \mu^2 d$ and $\Delta_2 = 1 + \mu^2 d$ we obtain $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} = \mu^2 d$ for the compression error of `HEAPRIX`. □

B 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, μ is compression of hash table, d is the dimension of the model, κ is condition number, ϵ is target accuracy, R is the number of communication rounds, and τ is the number of local updates. We start with the homogeneous setting comparison as follows:

Comparison of our results and existing ones for homogeneous and heterogeneous setting are given respectively Table 2 and Table 3

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:

Comparison with [35]. The reference [35] considers two-way compression from parameter server to 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 functions where ω^{Up} and ω^{Down} are uplink and downlink's compression noise (specializing to our case for the sake of comparison $\omega^{\text{Up}} = \omega^{\text{Down}} = \theta\left(d\right)$) for general heterogeneous data distribution. In contrast, while as pointed out in Remark 3.1 that 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 $1-\delta$.

C Theoretical Proofs

We will use the following fact (which is also used in [30, 15]) in proving results.

Fact 3 ([30, 15])**.** *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 5 ([14])**.** *The output of the compression operator $Q(\mathbf{x})$ is an unbiased estimator of its input $\mathbf{x}$, and its variance grows with the squared of the squared of ℓ_2-norm of its argument, i.e., $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}$ and $\mathbb{E}\left[\|Q(\mathbf{x}) - \mathbf{x}\|^2\right] \leq \omega\left\|\mathbf{x}\right\|^2$.*

We note that the sketching `PRIVIX` and `HEAPRIX`, satisfy Assumption 5 with $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively with probability $1-\delta$. Therefore, all the results in Theorem 1, are concluded with probability $1-\delta$ by plugging $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively into the corresponding convergence bounds.

C.1 Proof of Theorem 1

In this section, we study the convergence properties of our `FedSKETCH` method presented in Algorithm 4. Before stating the proofs for `FedSKETCH` in the homogeneous setting, we first mention the following intermediate lemmas.

Lemma 1. *Using unbiased compression and under Assumption 3, we have the following bound:*

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

Proof.

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

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

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

Next we show that from Assumptions 4, we have

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

To do so, note that

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

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 in i.i.d. manner at each local machine, and ③ immediately follows from Assumptions 3.

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

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

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

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

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:

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

$$\begin{aligned}
&\overset{①}{=}\frac{1}{2}\eta\sum_{c=0}^{\tau-1}\sum_{j=1}^{m} q_j\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + \|\nabla f(\boldsymbol{w}^{(r)}) - \nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2\right]\\
&\overset{②}{\leq}\frac{1}{2}\eta\sum_{c=0}^{\tau-1}\sum_{j=1}^{m} q_j\left[-\|\nabla f(\boldsymbol{w}^{(r)})\|_2^2 - \|\nabla f(\boldsymbol{w}_j^{(c,r)})\|_2^2 + L^2\|\boldsymbol{w}^{(r)} - \boldsymbol{w}_j^{(c,r)}\|_2^2\right] \qquad (11)
\end{aligned}$$

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 3 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

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

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 3. □

C.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 4 (non-convex)**.** *For* `FedSKETCH`(τ,η,γ)*, for all* $0 \leq t \leq R\tau - 1$*, under Assumptions 1 to 3, if the learning rate satisfies*

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

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:

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

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

Proof. Before proceeding to the proof of Theorem 4, 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 4 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:

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

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

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

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

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

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

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

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

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

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

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

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 convergence rate reduces to:*

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

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 $\tau = O\left(\frac{\omega+1}{k\epsilon}\right)$ local updates are necessary. We 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 convergence rate.

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

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

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:

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

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

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

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

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

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

which matches the rate obtained in [42]. 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 [42], which indicates the tightness of our analysis.

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

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

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

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

Proof. From (18) under condition:

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

we obtain:

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

which leads to the following bound:

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

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

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big]$$

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

□

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

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

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* $\tau = \frac{(\omega+1)}{k\left(\frac{\omega}{k}+1\right)\epsilon}$*. Additionally, we note that if* $\gamma \to \infty$*, yet* $R = O\left(\left(\frac{\omega}{k}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = \frac{(\omega+1)}{k\left(\frac{\omega}{k}+1\right)\epsilon}$ *will be necessary.*

C.1.3 Main result for the general convex setting

Theorem 6 (Convex)**.** *For a general convex function $f(\boldsymbol{w})$ with optimal solution $\boldsymbol{w}^{(*)}$, using* `FedSKETCH`(τ, η, γ) *to optimize $\tilde{f}(\boldsymbol{w}, \phi) = f(\boldsymbol{w}) + \frac{\phi}{2}\left\|\boldsymbol{w}\right\|^2$, for all $0 \leq t \leq R\tau - 1$, under Assumptions 1 to 3, if the learning rate satisfies*

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

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

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

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

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 5, we have:

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

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

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

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

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

thus the proof is complete. □

C.2 Proof of Theorem 2

The proof of Theorem 2 follows directly from the results in [14]. We first mention the general Theorem 7 from [14] for general compression noise ω. Next, since the sketching `PRIVIX` and `HEAPRIX`, satisfy Assumption 5 with $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively with probability $1-\delta$, all the results in Theorem 2, conclude from Theorem 7 with probability $1-\delta$ and plugging $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively into the corresponding convergence bounds. For the heterogeneous setting, the results in [14] requires the following extra assumption that naturally holds for the sketching:

Assumption 6 ([14])**.** *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 [14] here as follows:

Theorem 7. *Consider `FedCOMGATE` in [14]. If Assumptions 1, 4, 5 and 6 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* $\eta = \frac{1}{2L(\frac{\omega}{p}+1)\tau\gamma}$ *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* $R = O\left((\omega+1)\,\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ *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 5 with $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively with probablity $1-\delta$, we conclude the proofs of Theorem 2 using Theorem 7 with probability $1-\delta$ and plugging $\omega = \mu^2 d$ and $\omega = \mu^2 d - 1$ respectively into the convergence bounds. □

D Additional Numerical Experiments

D.1 Additional Plots for the MNIST Experiments

We are adding in this section, numerical runs for an intermediary number of local updates $\tau = 2$ and confirm for both cases the trend we observe in the plots of the main text regarding increasing this number. Our results illustrate the advantage of our proposed FS-HEAPRIX strategy in communication-efficient federated learning.

D.2 Homogeneous setting

In 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. The number of users is set to 50, and in each round of training we randomly pick half of the nodeds to be active (i.e. receiving data and performing local update). We can draw similar conclusion: FS-HEAPRIX consistently performs better than FS-PRIVIX, as well as SketchedSGD. The test accuracy increases with larger τ in homogeneous setting.

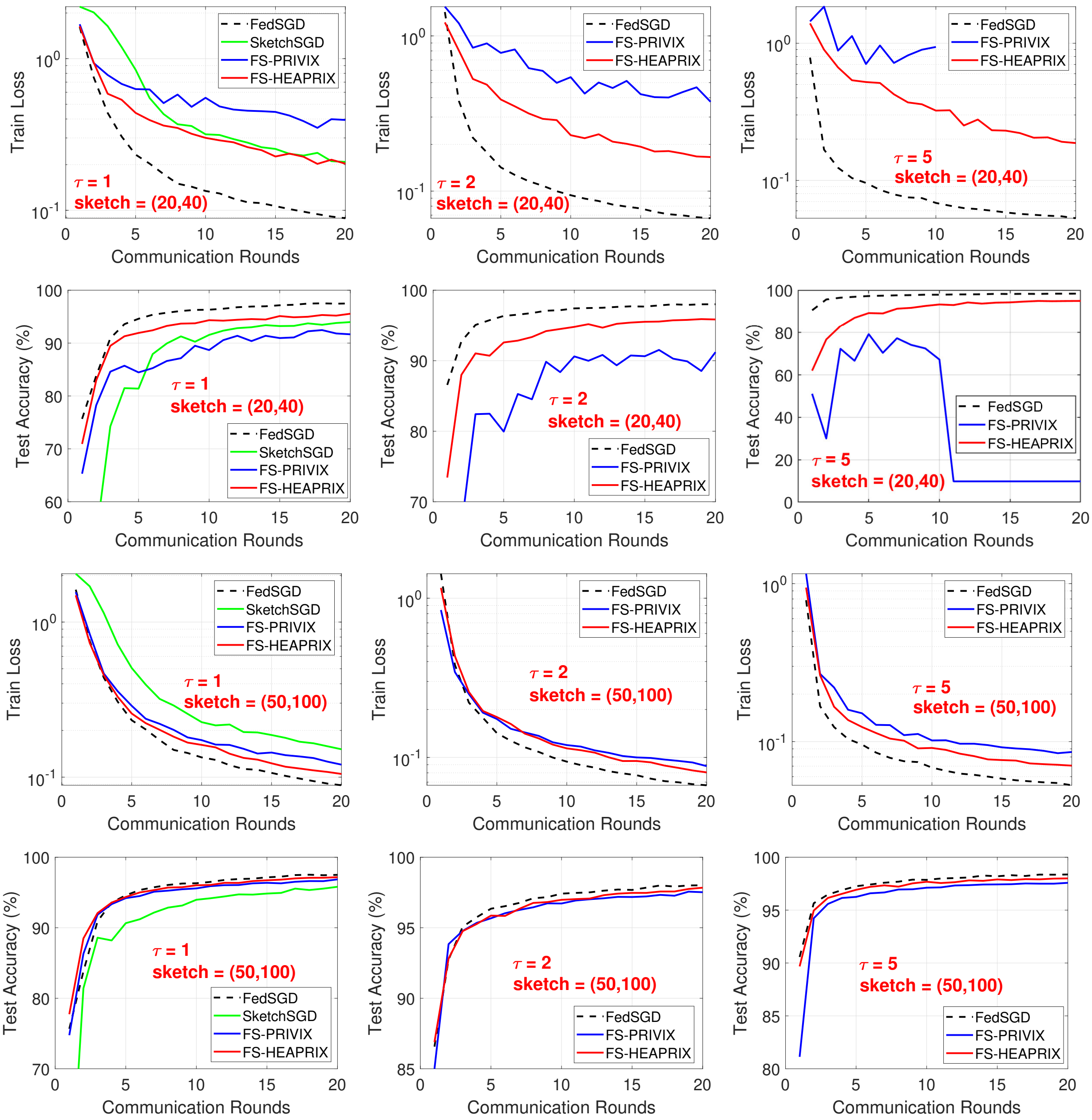


Figure 3 MNIST Homogeneous case: Comparison of compressed optimization methods on LeNet CNN architecture.

D.3 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 FS-PRIVIX method. SketchedSGD performs poorly in this case.

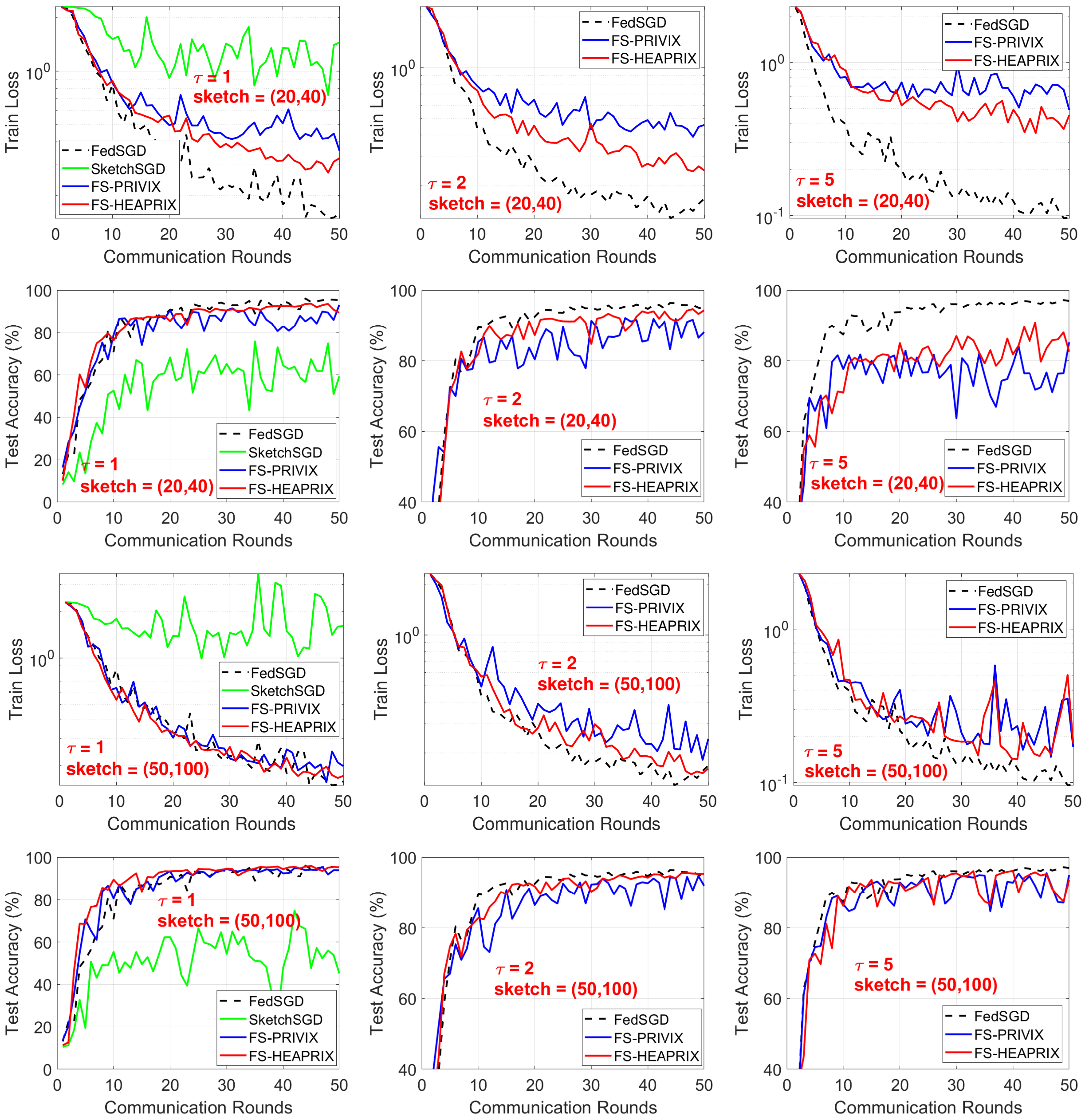
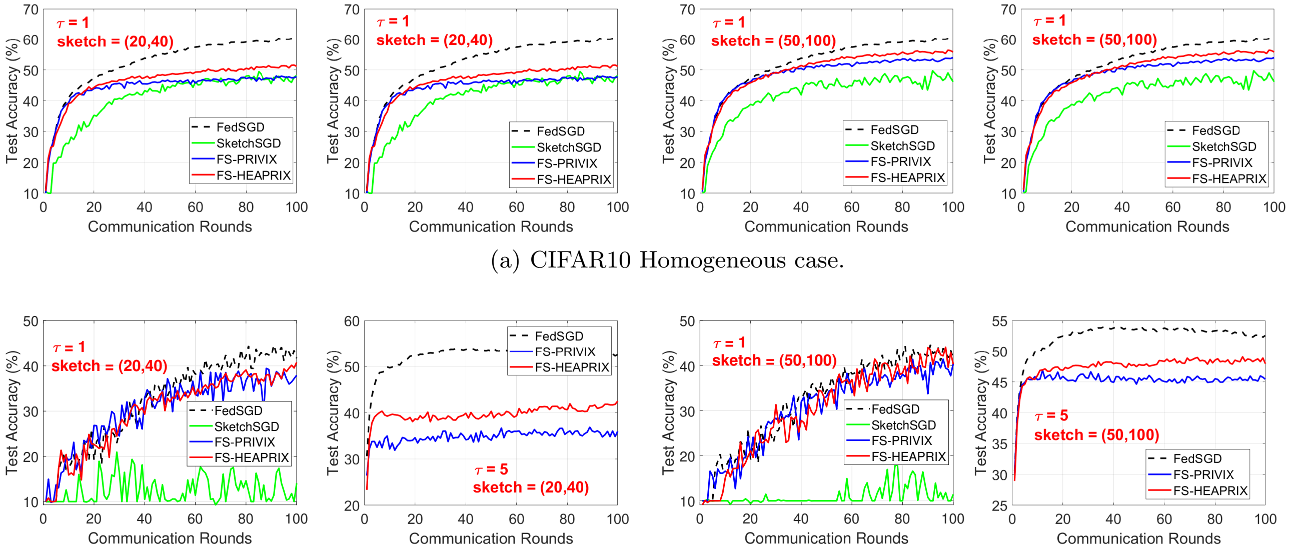


Figure 4 MNIST Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN architecture.

D.4 Additional Experiments: CIFAR-10



(a) CIFAR10 Homogeneous case.

(b) CIFAR10 Heterogeneous case.

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

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, for both 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.