Communication-Efficient Federated Learning via Sketching with Sharp Rates

Anonymous Authors[1]

Abstract

Communication complexity and data privacy are the two key challenges in Federated Learning (FL) where the goal is to perform a distributed learning through a large volume of devices. In this work, we introduce two new algorithms, namely `FedSKETCH` and `FedSKETCHGATE`, to address jointly both challenges and which are, respectively, intended to be used for homogeneous and heterogeneous data distribution settings. Our algorithms are based on a key and novel sketching technique, called `HEAPRIX` (`HP`) that is unbiased, compresses the accumulation of local gradients using count sketch, and exhibits communication-efficiency properties leveraging low-dimensional sketches. We provide sharp convergence guarantees of our algorithms and validate our theoretical findings with various sets of experiments.

1. Introduction

Federated Learning (FL) is an emerging framework for distributed large scale machine learning problems. In FL, data is distributed across devices (Konečnỳ et al., 2016; McMahan et al., 2017) and users are only allowed to communicate with the parameter server. Formally, the optimization problem across p distributed devices is defined as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d,\, \sum_{j=1}^{p} q_j=1} f(\boldsymbol{x}) \triangleq \sum_{j=1}^{p} q_j f_j(\boldsymbol{x}) \,, \tag{1}$$

where for device $j \in \{1,\dots,p\}$, $f_j(\boldsymbol{x}) = \mathbb{E}_{\xi\in\mathcal{D}_j}\left[L_j\left(\boldsymbol{x},\xi\right)\right]$, L_j is a loss function that measures the performance of model $\boldsymbol{x}$, ξ is a random variable distributed according to probability distribution $\mathcal{D}_j$, $q_j \triangleq \frac{n_j}{n}$ indicates the portion of data samples, n_j is the number of data shards and $n = \sum_{j=1}^{p} n_j$ is the total number of data samples. Note that contrary to the homogeneous setting where we assume $\{\mathcal{D}_j\}_{j=1}^{p}$ have the same distribution across devices and $L_i = L_j$, $1 \leq (i,j) \leq p$, in the heterogeneous setting these distributions and loss functions L_j can vary from a device to another.

There are several challenges that need to be addressed in FL in order to efficiently learn a global model that performs well in average for all devices:

– *Communication-efficiency*: There are often many devices communicating with the server, thus incurring immense communication overhead. One approach to reduce the number of communication rounds is using *local SGD with periodic averaging* (Zhou & Cong, 2018; Stich, 2019; Yu et al., 2019b; Wang & Joshi, 2018) which periodically averages local models after a few local updates, contrary to baseline SGD (Bottou & Bousquet, 2008) where gradient averaging is performed at each iteration. Local SGD has been proposed in (McMahan et al., 2017; Konečnỳ et al., 2016) under the FL setting and its convergence analysis is studied in (Stich, 2019; Wang & Joshi, 2018; Zhou & Cong, 2018; Yu et al., 2019b), later on improved in the followup references (Basu et al., 2019; Haddadpour & Mahdavi, 2019; Khaled et al., 2020; Stich & Karimireddy, 2019) for homogeneous setting. It is further extended to heterogeneous setting (Sahu et al., 2018; Haddadpour & Mahdavi, 2019; Karimireddy et al., 2019; Yu et al., 2019a; Li et al., 2020d; Liang et al., 2019). The second approach dealing with communication cost aims at reducing the size of communicated message per communication round, such as gradient quantization (Alistarh et al., 2017; Bernstein et al., 2018; Tang et al., 2018; Wen et al., 2017; Wu et al., 2018) or sparsification (Stich et al., 2018; Alistarh et al., 2018; Lin et al., 2018; Stich & Karimireddy, 2019).

–*Data heterogeneity*: Since locally generated data in each device may come from different distribution, local computations involved in FL setting can lead to poor convergence error in practice (Li et al., 2020a; Liang et al., 2019). To mitigate the negative impact of data heterogeneity, (Horváth et al., 2019; Liang et al., 2019; Karimireddy et al., 2019; Haddadpour et al., 2020) suggest applying variance reduction or gradient tracking techniques along local computations.

–*Privacy* (Geyer et al., 2017; Hardy et al., 2017): Privacy has been widely addressed by injecting an additional layer of randomness to respect differential-privacy property (McMa-

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

han et al., 2018) or using cryptography-based approaches under secure multi-party computation (Bonawitz et al., 2017). Further study related to FL setting can be found in recent surveys (Li et al., 2020a) and (Kairouz et al., 2019).

To jointly tackle the aforementioned challenges in FL, sketching based algorithms (Charikar et al., 2004; Cormode & Muthukrishnan, 2005; Kleinberg, 2003; Li et al., 2008) are promising methods. For instance, to reduce the communication cost, (Ivkin et al., 2019) develops a distributed SGD algorithm using sketching, provides its convergence analysis in the homogeneous setting, and establishes a communication complexity of order $\mathcal{O}(\log(d))$ per round, where d is the dimension of the gradient vector compared to $\mathcal{O}(d)$ complexity per round of baseline mini-batch SGD. Nonetheless, the proposed sketching scheme in (Ivkin et al., 2019), built from a communication-efficiency perspective, is based on a deterministic procedure which requires access to the exact information of the gradients, thus not meeting the privacy-preserving criteria. This systemic issue is partially addressed in (Rothchild et al., 2020).

Focusing on privacy, (Li et al., 2019) derives a single framework in order to address these issues and introduces `DiffSketch` algorithm, based on the Count Sketch operator, yet does not provide its convergence analysis. Besides, the estimation error of `DiffSketch` is higher than the sketching scheme in (Ivkin et al., 2019) which could lead to poor convergence.

In this paper, we propose new methods to tackle the communication bottleneck of FL baselines. Our main contributions are summarized as follows:

- We provide a new algorithm – `HEAPRIX` (`HP`) – and theoretically show that it reduces the cost of communication, based on unbiased sketching without requiring the broadcast of exact values of gradients to the server. Based on `HP`, we develop general algorithms for communication-efficient and sketch-based FL, namely `FedSKETCH` and `FedSKETCHGATE` for homogeneous and heterogeneous data distribution settings respectively.
- We establish non-asymptotic convergence bounds for Polyak-Łojasiewicz (PL), convex and non-convex functions in Theorems 1 and 2 in both homogeneous and heterogeneous cases, and highlight an improvement in the number of iterations to reach a stationary point. We also provide *sharper* convergence analysis for the `PRIVIX(PR)`/`DiffSketch`[1] algorithm proposed in (Li et al., 2019).
- We illustrate the benefits of `FedSKETCH` and `FedSKETCHGATE` over baseline methods through a number of experiments. The latter shows the advantages of the `HP` compression method achieving comparable test accuracy as Federated SGD (`FedSGD`) while compressing the information exchanged between devices and server.

Notation: We denote the number of communication rounds and bits per round and per device by R and B respectively. The count sketch of vector $\boldsymbol{x}$ is designated by $\mathbf{S}(\boldsymbol{x})$. $[p]$ denotes the set $\{1,\dots,p\}$.

2. Compression using Count Sketch

Throughout the paper, we employ the commonly used `Count Sketch` (Charikar et al., 2004) as building component of our algorithms. Please refer to the Appendix for the detailed Count Sketch algorithm.

There are various types of sketching algorithms which are developed based on count sketching that we develop in the following subsections.

2.1. Sketching based Unbiased Compressor

We define an unbiased compressor as follows:

Definition 1 (Unbiased compressor)**.** *We call the randomized function $\mathcal{C}:\mathbb{R}^d\to\mathbb{R}^d$ an unbiased compression operator with $\Delta\geq 1$, if*

$$\mathbb{E}\left[\mathcal{C}(\boldsymbol{x})\right]=\boldsymbol{x}\quad \textit{and}\quad \mathbb{E}\left[\left\|\mathcal{C}(\boldsymbol{x})\right\|_2^2\right]\leq\Delta\left\|\boldsymbol{x}\right\|_2^2 .$$

We denote this class of compressors by $\mathbb{U}(\Delta)$.

This definition leads to the following property

$$\mathbb{E}\left[\left\|\mathcal{C}(\boldsymbol{x})-\boldsymbol{x}\right\|_2^2\right]\leq(\Delta-1)\left\|\boldsymbol{x}\right\|_2^2 .$$

This property allows us to control the noise of the compression. Note that if we let $\Delta=1$, then our algorithm reduces to the case of no compression.

For instance, `PR` is an unbiased compressor which obtains an estimate of input $\boldsymbol{x}$ from a count sketch noted $\boldsymbol{S}(\boldsymbol{x})$. For more detail please see (Li et al., 2019), or Algorithm 6 in the Appendix. We give below a useful property of Count Sketch for our theoretical analysis.

Property 1 ((Li et al., 2019))**.** *For any $\boldsymbol{x}\in\mathbb{R}^d$:*

1) Unbiased estimation: As in (Li et al., 2019), we have $\mathbb{E}_{\mathbf{S}}\left[\mathtt{PR}\left[\mathbf{S}\left(\boldsymbol{x}\right)\right]\right]=\boldsymbol{x}$.

2)Bounded variance: For the given $m<d$, $t=\mathcal{O}\left(\ln\left(\frac{d}{\delta}\right)\right)$ with probability $1-\delta$ we have:

$$\mathbb{E}_{\mathbf{S}}\left[\left\|\mathtt{PR}\left[\mathbf{S}\left(\boldsymbol{x}\right)\right]-\boldsymbol{x}\right\|_2^2\right]\leq c\frac{d}{m}\left\|\boldsymbol{x}\right\|_2^2 ,$$

where c ($e\leq c<m$) is a positive constant independent of the dimension of the input, d.

[1] We use `PRIVIX` (`PR`) and `DiffSketch` (Li et al., 2019) interchangeably throughout the paper.

We note that this bounded variance assumption does not necessary mean that compression is happening since dimension d may be relatively large. Thus, with probability $1-\delta$, we obtain $\texttt{PR} \in \mathbb{U}(1+c\frac{d}{m})$. $\Delta = 1 + c\frac{d}{m}$ implies that if $m \to d$, then $\Delta \to 1 + c$, indicates a noisy reconstruction. (Li et al., 2019) shows that if the data is normally distributed, PR is differentially private (Dwork, 2006), up to additional assumptions and algorithmic design choices.

2.2. Sketching based Biased Compressor

A biased compressor is defined as follows:

Definition 2 (Biased compressor)**.** *A (randomized) function, $\mathcal{C}: \mathbb{R}^d \to \mathbb{R}^d$ belongs to $\mathbb{C}(\Delta, \alpha)$, a class of compression operators with $\alpha > 0$ and $\Delta \geq 1$, if*

$$\mathbb{E}\left[\|\alpha\boldsymbol{x} - \mathcal{C}(\boldsymbol{x})\|_2^2\right] \leq \left(1 - \frac{1}{\Delta}\right)\|\boldsymbol{x}\|_2^2 .$$

It is proven in (Horváth & Richtárik, 2020) that $\mathbb{U}(\Delta) \subset \mathbb{C}(\Delta, \alpha)$. An example of biased compression using sketching methods and using a top_m operator is provided below:

Algorithm 1 HEAVYMIX (HX) (Modified (Ivkin et al., 2019))

```
Inputs: $\mathbf{S}(\mathbf{g})$; parameter $m$
Query the vector $\tilde{\mathbf{g}} \in \mathbb{R}^d$ from $\mathbf{S}(\mathbf{g})$:
Query $\hat{\ell}_2^2 = (1 \pm 0.5)\|\mathbf{g}\|^2$ from sketch $\mathbf{S}(\mathbf{g})$
$\forall j$ query $\hat{\mathbf{g}}_j^2 = \hat{\mathbf{g}}_j^2 \pm \frac{1}{2m}\|\mathbf{g}\|^2$ from sketch $\mathbf{S}(\mathbf{g})$
$H = \{j | \hat{\mathbf{g}}_j \geq \frac{\hat{\ell}_2^2}{m}\}$ and $NH = \{j | \hat{\mathbf{g}}_j < \frac{\hat{\ell}_2^2}{m}\}$
$\text{Top}_m = H \cup \text{rand}_\ell(NH)$, where $\ell = m - |H|$
Get exact values of $\text{Top}_m$
Output: $\tilde{\mathbf{g}} : \forall j \in \text{Top}_m : \tilde{\mathbf{g}}_i = \mathbf{g}_i$ else $\mathbf{g}_i = 0$
```

Following (Ivkin et al., 2019), HEAVYMIX (HX) with sketch size $\Theta\left(m \log\left(\frac{d}{\delta}\right)\right)$ is a biased compressor with $\alpha = 1$ and $\Delta = d/m$ with probability $\geq 1 - \delta$. In other words, with probability $1 - \delta$, $\texttt{HX} \in \mathcal{C}(\frac{d}{m}, 1)$. Note that Algorithm 1 is a variant of the sketching algorithm developed in (Ivkin et al., 2019) with the distinction that HX does not require any second round of communication to obtain the exact values of top_m. This is mainly because in SKETCHED-SGD (Ivkin et al., 2019), the server receives the exact values of *the average of the sketches* while HX obtains exact local values. Additionally, while HX has a smaller estimation error compared to PR, in PR the central server does need to have access to the exact values of local gradient providing user privacy as underlined in (Li et al., 2019).

In the following, we introduce HX which is built upon HX and PR methods.

2.3. Sketching based Induced Compressor

From Theorem 3 in (Horváth & Richtárik, 2020), stating that a biased compressor can be converted into an unbiased one such that, for $\mathcal{C}_1 \in \mathbb{C}(\Delta_1)$ with $\alpha = 1$, if one chooses $\mathcal{C}_2 \in \mathbb{U}(\Delta_2)$, then the induced compressor $\mathcal{C}: x \mapsto \mathcal{C}_1(\mathbf{x}) + \mathcal{C}_2(\mathbf{x} - \mathcal{C}_1(\mathbf{x}))$ belongs to $\mathbb{U}(\Delta)$ with $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1}$.

Algorithm 2 HP

```
Inputs: $\boldsymbol{x} \in \mathbb{R}^d, \mathbf{S}_{m\times t}, m < t$
Approximate $\mathbf{S}(\boldsymbol{x})$ using HX
Approximate $\mathbf{S}(\boldsymbol{x} - \texttt{HX}[\mathbf{S}(\boldsymbol{x})])$ with PR
Output:
    $\Phi(\boldsymbol{x}) \triangleq \texttt{HX}[\mathbf{S}(\boldsymbol{x})] + \texttt{PR}[\mathbf{S}(\boldsymbol{x} - \texttt{HX}[\mathbf{S}(\boldsymbol{x})])]$ .
```

Based on this notion, Algorithm 2 proposes an induced sketching algorithm by utilizing HX and PR for $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively where the reconstruction of input $\mathbf{x}$ is performed using hash table $\mathbf{S}$ and $\mathbf{x}$, similar to PR and HX. Note that if $m \to d$, then $\mathcal{C}(\boldsymbol{x}) \to \boldsymbol{x}$, implying that the convergence rate can be improved by decreasing the size of compression m.

Corollary 1. *Based on Theorem 3 of (Horváth & Richtárik, 2020), HX in Algorithm 2 satisfies $\mathcal{C}(\boldsymbol{x}) \in \mathbb{U}(c\frac{d}{m})$.*

3. FedSKETCH and FedSKETCHGATE

We introduce two new algorithms for both homogeneous and heterogeneous settings.

3.1. Homogeneous Setting

In FedSKETCH, the number of local updates, between two consecutive communication rounds, at device j is denoted by τ. Unlike (Haddadpour et al., 2020), the server does not store any global model, rather, device j has two models: $\boldsymbol{x}^{(r)}$ and $\boldsymbol{x}_j^{(\ell,r)}$, which are respectively the global and local models. We develop FedSKETCH in Algorithm 3 with a variant of this algorithm implementing HP. For this variant, we need to have an additional communication round between the server and worker j to aggregate $\delta_j^{(r)} \triangleq \mathbf{S}_j\left[\texttt{HX}(\mathbf{S}^{(r)})\right]$ (Lines 3 and 3) to compute $\mathbf{S}^{(r)} = \frac{1}{k}\sum_{j\in\mathcal{K}}\mathbf{S}_j^{(r)}$. The main difference between FedSKETCH and DiffSketch in (Li et al., 2019) is that we use distinct local and global learning rates. Furthermore, unlike (Li et al., 2019), we do not add local Gaussian noise *as privacy is not the main focus of this paper*.

Algorithmic comparison with (Haddadpour et al., 2020) An important feature of our algorithm is that due to a lower dimension of the count sketch, the resulting averaged quantities received by the server are also of lower dimension. Therefore, our algorithms exploit a bidirectional compression during the communication phases between server and

Algorithm 3 FedSKETCH(R, τ, η, γ)

```
Inputs: $\boldsymbol{x}^{(0)}$: initial model shared by local devices, global and local learning rates $\gamma$ and $\eta$, respectively
for $r = 0, \ldots, R-1$ do
parallel for device $j \in \mathcal{K}^{(r)}$ do:
    if PRIVIX variant:
        $$\boldsymbol{\Phi}^{(r)} \triangleq \texttt{PR}\left[\mathbf{S}^{(r-1)}\right]$$
    if HEAPRIX variant:
        $$\boldsymbol{\Phi}^{(r)} \triangleq \texttt{HX}\left[\mathbf{S}^{(r-1)}\right] + \texttt{PR}\left[\mathbf{S}^{(r-1)} - \tilde{\mathbf{S}}^{(r-1)}\right]$$
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma\boldsymbol{\Phi}^{(r)}$ and $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $\ell = 0, \ldots, \tau-1$ do
        Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)}$
        Update $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta\, \tilde{\mathbf{g}}_j^{(\ell,r)}$
    end for
    Device $j$ broadcasts $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$.
    Server computes $\mathbf{S}^{(r)} = \frac{1}{k}\sum_{j\in\mathcal{K}} \mathbf{S}_j^{(r)}$ .
    Server broadcasts $\mathbf{S}^{(r)}$ to devices in randomly drawn devices $\mathcal{K}^{(r)}$.
    if HEAPRIX variant:
        Second round of communication for computing $\delta_j^{(r)} := \mathbf{S}_j\left[\texttt{HX}(\mathbf{S}^{(r)})\right]$ and broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{k}\sum_{j\in\mathcal{K}} \delta_j^{(r)}$ to devices in set $\mathcal{K}^{(r)}$
end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$
```

devices. As a result, for the case of large quantization error $\omega = \theta(\frac{d}{m})$ as shown in (Haddadpour et al., 2020), our algorithms can outperform those developed in (Haddadpour et al., 2020) if sufficiently large hash tables are used and the uplink communication cost is high. Furthermore, while in (Haddadpour et al., 2020), server stores a global model and aggregates the partial gradients from devices which can enable the server to extract some information regarding the device's data, in contrast, in our algorithms server does not store the global model and only broadcasts the average sketches.

Remark 1. *As discussed in (Horváth & Richtárik, 2020), while induced compressors transform a biased compressor into an unbiased one at the cost of doubling the communication cost since the devices need to send $\mathcal{C}_1(\boldsymbol{x})$ and $\mathcal{C}_2\left(\boldsymbol{x} - \mathcal{C}_1\left(\boldsymbol{x}\right)\right)$ separately, we emphasize that with* HP*, due to the use of sketching, the extra communication round cost is compensated with lower number of bits per round thanks to the lower dimension of the sketches.*

Algorithm 4 FedSKETCHGATE(R, τ, η, γ)

```
Inputs: $\boldsymbol{x}^{(0)} = \boldsymbol{x}_j^{(0)}$ shared by all local devices, global and local learning rates $\gamma$ and $\eta$.
for $r = 0, \ldots, R-1$ do
parallel for device $j = 1, \ldots, p$ do:
    if PRIVIX variant:
        $$\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau}\left[\texttt{PR}\left(\mathbf{S}^{(r-1)}\right) - \texttt{PR}\left(\mathbf{S}_j^{(r-1)}\right)\right]$$
      where $\boldsymbol{\Phi}^{(r)} \triangleq \texttt{PR}(\mathbf{S}^{(r-1)})$
    if HEAPRIX variant:
        $$\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau}\left(\boldsymbol{\Phi}^{(r)} - \boldsymbol{\Phi}_j^{(r)}\right)$$
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma\boldsymbol{\Phi}^{(r)}$ and $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $\ell = 0, \ldots, \tau-1$ do
        Sample mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)}$
        $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta\left(\tilde{\mathbf{g}}_j^{(\ell,r)} - \mathbf{c}_j^{(r)}\right)$
    end for
    Device $j$ broadcasts $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$.
    Server computes $\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1} \mathbf{S}_j^{(r)}$ and broadcasts $\mathbf{S}^{(r)}$ to all devices.
    if HEAPRIX variant:
        Device $j$ computes $\boldsymbol{\Phi}_j^{(r)} \triangleq \texttt{HP}[\mathbf{S}_j^{(r)}]$.
        Second round of communication to obtain $\delta_j^{(r)} := \mathbf{S}_j\left(\texttt{HX}[\mathbf{S}^{(r)}]\right)$.
        Broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{p}\sum_{j=1}^{p} \delta_j^{(r)}$ to devices.
end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$
```

Benefits of HEAPRIX (HP) based algorithms: Corollary 1 states that, unlike PR, HP compression noise can be made as small as possible using larger hash size. In the distributed setting, contrary to SKETCHED-SGD (Ivkin et al., 2019) where decompressing is happening in the server, HP does not require to have access to exact top$_m$ values of the input rather it can only have access to sketches of aggregated local gradients. This is because HP uses HX *where decompressing is performed at each device locally, thus not requiring server to have exact values of gradients of each device*. In other words, HP-based algorithm leverages the best of both: the *unbiasedness* of PR while using *heavy hitters* as in HX.

3.2. Heterogeneous Setting

In this section, we focus on the optimization problem of (1) where $q_1 = \ldots = q_p = \frac{1}{p}$ with full device participation ($k = p$). These results can be extended to the scenario with devices sampling. For non i.i.d. data, the FedSKETCH algorithm, designed for homogeneous setting, may fail to perform well in practice. The main reason is that in FL, devices are using local stochastic descent direction which could be different than global descent direction when the data distribution are non-identical. For that reason, to mitigate the negative impact of data heterogeneity, we introduce a new algorithm called FedSKETCHGATE described in Algorithm 4. The main idea is that using an approximation of the global gradient, $\mathbf{c}_j^{(r)}$ allows to correct the local gradient direction, see gradient tracking technique applied in (Liang et al., 2019; Haddadpour et al., 2020). Using PR variant, the correction vector $\mathbf{c}_j^{(r)}$ at device j and communication round r is computed in Line 4. Using HP variant, FedSKETCHGATE updates $\tilde{\mathbf{S}}^{(r)}$ via Line 4.

Remark 2. *Most of the existing communication-efficient algorithms with compression only consider gradient-compression from devices to server. However, Algorithms 3 and 4 improve the communication efficiency from server to devices as it exploits low-dimensional sketches in a bidirectional manner.*

For both FedSKETCH and FedSKETCHGATE, HP variant requires a second round of communication, unlike PR. Therefore, in Cross-Device FL setting, where there could be millions of devices, HP variant may not be practical, and we note that it could be more suitable for Cross-Silo FL setting.

4. Convergence Analysis

We first state commonly used assumptions required in the following convergence analysis (reminder of our notations can be found Table 1 of the Appendix).

Assumption 1 (Smoothness and Lower Boundedness)**.** *The local objective function $f_j(\cdot)$ of device j is differentiable for $j \in [p]$ and L-smooth, i.e., $\|\nabla f_j(\boldsymbol{x}) - \nabla f_j(\mathbf{y})\| \leq L\|\boldsymbol{x} - \mathbf{y}\|$, $\forall\, \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d$. Moreover, the optimal objective function $f(\cdot)$ is bounded below by $f^* := \min_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$.*

We present our results for PL, convex and general non-convex objectives. (Karimi et al., 2016) show that strong convexity implies PL condition with same module (PL objectives can also be non-convex, hence PL condition does not imply the strong convexity necessarily).

4.1. Convergence of FEDSKETCH

We start with the homogeneous case where data is i.i.d. among local devices, and therefore, the stochastic local gradient of each worker is an unbiased estimator of the global gradient. We make the following assumption under that setting:

Assumption 2 (Bounded Variance)**.** *For all $j \in [m]$, we can sample an independent mini-batch ℓ_j of size $|\xi_j^{(\ell,r)}| = b$ and compute an unbiased stochastic gradient $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{x}; \xi_j)$, $\mathbb{E}_{\xi_j}[\tilde{\mathbf{g}}_j] = \nabla f(\boldsymbol{x}) = \mathbf{g}$ with the variance bounded by a constant σ^2, i.e., $\mathbb{E}_{\xi_j}\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}\|^2\right] \leq \sigma^2$.*

Theorem 1. *Suppose Assumptions 1-2 holds. 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 $1-\delta$ we have:*

*In the **non-convex** case, $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ satisfies $\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2\right] \leq \epsilon$ if:*

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

*In the **PL or strongly convex** case, $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ satisfies $\mathbb{E}[f(\boldsymbol{x}^{(R-1)}) - f(\boldsymbol{x}^{(*)})] \leq \epsilon$ if we set:*

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

where constant c comes from property 1. The bounds in Theorem 1 suggest that, under the homogeneous setting, if we set $d = m$, i.e. no compression, the number of communication rounds to achieve the ϵ-error matches with the number of iterations required to achieve the same error under a centralized setting. Furthermore, for non-convex objective, FS-HEAPRIX improves the computational complexity over FS-PRIVIX, and for PL objectives FS-HEAPRIX improves communication complexity over FS-PRIVIX. Furthermore, this improvement is validated through our experiments in Section 5. Additionally, we can see that the computational complexity scales down partially with the number of sampled devices. To further stress on the impact of using sketching methods, we also compare our results with prior works in terms of total number of communicated bits per device. Please refer to Section 3 of the Appendix for the convergence bounds using convex objectives.

Comparison with (Ivkin et al., 2019) From a privacy aspect, we note that (Ivkin et al., 2019) requires for the central server to have access to exact values of top_m gradients, hence does not preserve privacy, whereas our schemes do not need those exact values. From a communication cost point of view, for strongly convex objective and compared to (Ivkin et al., 2019), we improve the total communication per worker from $RB = O\left(\frac{d}{\epsilon}\log\left(\frac{d}{\delta\sqrt{\epsilon}}\max\left(\frac{d}{m}, \frac{1}{\sqrt{\epsilon}}\right)\right)\right)$ to

$$RB = O\left(\kappa(cd - m + \tfrac{m}{k})\log\tfrac{1}{\epsilon}\log\left(\tfrac{\kappa d}{\delta}(\tfrac{cd-m}{m} + 1/k)\log\tfrac{1}{\epsilon}\right)\right).$$

We note that while reducing communication cost, our scheme requires $\tau = O(cd/m((\frac{cd-m}{m} + 1/k)\epsilon)) > 1$, which scales down with the number of sampled devices k. Moreover, unlike (Ivkin et al., 2019), we do not use the classical bounded gradient assumption and thus obtain stronger results with weaker assumptions. Regarding general non-convex objectives, our result improves the total communication cost per worker displayed in (Ivkin et al., 2019) from $RB = O\left(\max(\frac{1}{\epsilon^2}, \frac{d^2}{k^2\epsilon}\right) \log(\frac{d}{\delta}\max(\frac{1}{\epsilon^2}, \frac{d^2}{k^2\epsilon})))$ for *only one device* to $RB = O(\frac{m}{\epsilon}\log(\frac{d}{\epsilon\delta}))$. We also highlight that we can obtain similar rates for Algorithm 3 in heterogeneous environment if we make the additional uniform boundedness gradient assumption.

Note: Such improved communication cost over prior works can be explained by – the joint exploitation of *sketching* – the reduction of the dimension of communicated messages – the use of *local updates* – the reduction of the number of communication rounds reaching a specific convergence error.

4.2. Convergence of FedSKETCHGATE

We start with a bounded local variance assumption:

Assumption 3 (Bounded Local Variance). *For all $j \in [p]$, we can sample an independent mini-batch Ξ_j of size $|\xi_j| = b$ and compute an unbiased stochastic gradient $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{x}; \xi_j)$ with $\mathbb{E}_\xi[\tilde{\mathbf{g}}_j] = \nabla f_j(\boldsymbol{x}) = \mathbf{g}_j$. Moreover, the variance of local stochastic gradients is bounded such that $\mathbb{E}_\xi\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}_j\|^2\right] \leq \sigma^2$.*

Theorem 2. *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 $1-\delta$ we have*

*In the **non-convex** case, $\eta = \frac{1}{L\gamma}\sqrt{\frac{mp}{R\tau(cd)}}$, $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ satisfies $\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2\right] \leq \epsilon$ if:*

- *FS-PRIVIX:*

$$R = O((cd+m)/m\epsilon) \quad \text{and} \quad \tau = O(1/(p\epsilon))\ .$$

- *FS-HEAPRIX: $R = O(d/m\epsilon)$ and $\tau = O(1/(p\epsilon))$.*

*In the **PL or Strongly convex** case, $\{\boldsymbol{x}^{(r)}\}_{r=>0}$ satisfies $\mathbb{E}\Big[f(\boldsymbol{x}^{(R-1)}) - f(\boldsymbol{x}^{(*)})\Big] \leq \epsilon$ if:*

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

Theorem 2 implies that the number of communication rounds and local updates are similar to the corresponding quantities in homogeneous setting except for the non-convex case where the number of rounds also depends on the compression rate (summarized Table 2-3 of the Appendix). For the convergence result of convex objectives please see Section 3 in appendix.

We note that the convergence analysis of FS-PRIVIX provided in (Li et al., 2019) for convex objectives is further tightened in our contribution. Moreover, FS-HEAPRIX improves the communication complexity of FS-PRIVIX for both PL and non-convex objectives which is empirically validated in Figures 1 and 2.

4.3. Comparison with Prior Methods

Main competing baselines of our methods are distributed algorithms based on sketching. Nonetheless, we also compare with prior non-sketching based distributed algorithms ((Karimireddy et al., 2019; Basu et al., 2019; Reisizadeh et al., 2020; Haddadpour et al., 2020)) in Section C of the Appendix.

(Li et al., 2019). Note that our convergence analysis does not rely on the bounded gradient assumption. We also improve both the number of communication rounds R and the size of transmitted bits B per communication round (please see Table 3 of Section C in appendix). Additionally, we highlight that, while (Li et al., 2019) provides a convergence analysis for convex objectives, our analysis holds for PL (thus strongly convex case), general convex and general non-convex objectives.

(Rothchild et al., 2020). Due to gradient tracking, our algorithm tackles data heterogeneity, while algorithms in (Rothchild et al., 2020) do not. Thereby, in FedSKETCHGATE each device has to store an additional state vector compared to (Rothchild et al., 2020). Yet, as our method is built upon an unbiased compressor, server does not need to store any additional error correction vector. The convergence results for both FetchSGD variants in (Rothchild et al., 2020) rely on the uniform bounded gradient assumption which may not be applicable with L-smoothness assumption when data distribution is highly heterogeneous, as it is commonly the case in FL, see (Khaled et al., 2020). Besides, Theorem 1 (Rothchild et al., 2020) assumes that *Contraction Holds* for the sequence of gradients which may not hold in practice, yet based on this strong assumption, their total communication cost (RB) in order to achieve ϵ error is $RB = O\left(m\max(\frac{1}{\epsilon^2}, \frac{d^2-dm}{m^2\epsilon})\log\left(\frac{d}{\delta}\max(\frac{1}{\epsilon^2}, \frac{d^2-dm}{m^2\epsilon})\right)\right)$. For the sake of comparison, we let the compression ratio in (Rothchild et al., 2020) to be $\frac{m}{cd}$. In contrast, without any extra assumptions, our results in Theorem 2 for PRIVIX and HEAPRIX are respectively $RB = O(\frac{(cd+m)}{\epsilon}\log(\frac{(\frac{cd^2}{m})+d}{\epsilon\delta}))$ and $RB = O(c\frac{d}{\epsilon}\log(\frac{cd^2}{\epsilon m\delta}))$ which improves the total communication cost of Theorem 1 in (Rothchild et al., 2020) under regimes such that $\frac{1}{\epsilon} \geq d$ or $d \gg m$. Theorem 2 in (Rothchild et al., 2020) is based the *Sliding Window Heavy Hitters* assumption, which is similar to the gradient diversity assumption in (Li et al., 2020c; Haddadpour & Mahdavi, 2019). Under that assumption, the total communication cost is shown to be $RB = O\left(\frac{m \max(I^{2/3}, 2-\alpha)}{\epsilon^3 \alpha} \log\left(\frac{d \max(I^{2/3}, 2-\alpha)}{\epsilon^3 \delta}\right)\right)$ where I is a constant related to the window of gradients. We improve this bound under weaker assumptions in a regime where $\frac{I^{2/3}}{\epsilon^2} \geq d$. We also provide bounds for PL, convex and non-convex objectives contrary to (Rothchild et al., 2020).

Comparison with SCAFFOLD. To compare with (Karimireddy et al., 2019) which does not use gradient compression, we let $m = d$ (no compression). In this case, similar to (Haddadpour et al., 2020), our communication complexities and number of local updates match with corresponding bounds obtained by SCAFFOLD with difference that in downlink (from devices to server) we only send one vector while SCAFFOLD needs to send two vectors (additional control variate). Extensive comparison with related methods, involving sketches or not, such as (Rothchild et al., 2020) or (Karimireddy et al., 2019), can be found Section C of the Appendix. Additionally, unlike (Rothchild et al., 2020) only focusing on non-convex objectives, we provide the convergence analysis for PL (thus strongly convex case), general convex and general non-convex objectives. Finally, the algorithms in (Rothchild et al., 2020) require additional server memory to store the compression error correction while our does not.

5. Numerical Study

In this section, we provide empirical results on MNIST benchmark dataset to demonstrate the effectiveness of our proposed algorithms. We train LeNet-5 Convolutional Neural Network (CNN) architecture introduced in (LeCun et al., 1998), with 60 000 parameters. We compare Federated SGD (`FedSGD`) as the full-precision baseline, along with four sketching methods `SketchSGD` (Ivkin et al., 2019), `FetchSGD` (Rothchild et al., 2020), and two FedSketch variants `FS-PRIVIX` and `FS-HEAPRIX`. Note that in Algorithm 3, `FS-PRIVIX` with global learning rate $\gamma = 1$ is equivalent to the `DiffSketch` algorithm proposed in (Li et al., 2020c). Also, `SketchSGD` is slightly modified to compress the change in local weights (instead of local gradient in every iteration), and `FetchSGD` is implemented with second round of communication for fairness. (The original proposal does not include second round of communication, which performs worse with small sketch size.) As suggested in (Rothchild et al., 2020), the momentum factor of `FetchSGD` is set to 0.9, and we also follow some recommended implementation tricks to improve its performance, which are detailed in the Appendix. The number of workers is set to 50 and we report the results for 1 and 5 local epochs. A local epoch is finished when all workers go through their local data samples once. The local batch size is 30. In each round, we randomly choose half of the devices to be active. We tune the learning rates (η and γ, if applicable) over log-scale and report the best results, for both *homogeneous* and *heterogeneous* setting. In the former case, each device receives uniformly drawn data samples, and in the latter, it only receives samples from one or two classes among ten.

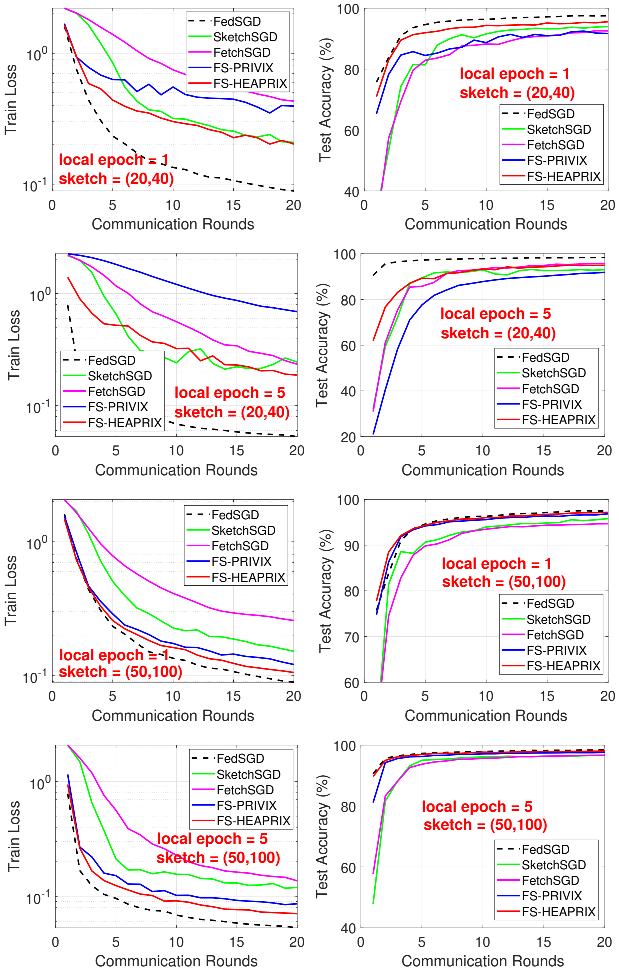

Figure 1. Homogeneous case: Comparison of compressed optimization methods on LeNet CNN.

Homogeneous case. In Figure 1, we provide the training loss and test accuracy with different number of local epochs and sketch size, $(t, k) = (20, 40)$ and $(50, 100)$. Note that, these two choices of sketch size correspond to a $75\times$ and $12\times$ compression ratio, respectively. We conclude that

- In general, increasing the compression ratio sacrifices the learning performance. In all cases, `FS-HEAPRIX` performs the best in terms of both training objective

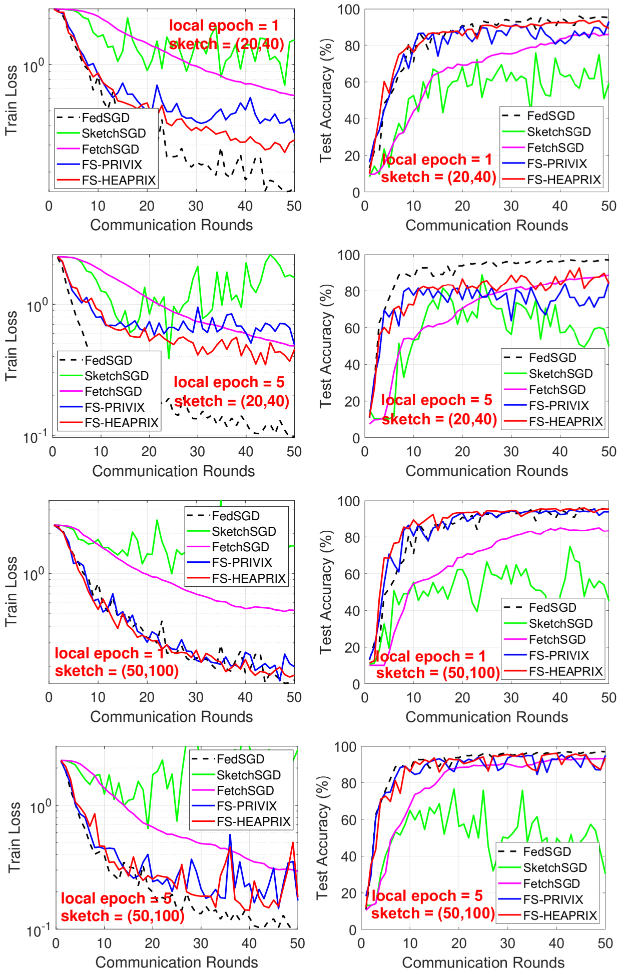

Figure 2. Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN.

and test accuracy, among all compressed methods.

- FS−HEAPRIX is better than FS−PRIVIX, especially with small sketches (high compression ratio). FS−HEAPRIX yields acceptable extra test error compared to full-precision FedSGD, particularly when considering the high compression ratio (e.g., 75×).

- The training performance of FS−HEAPRIX improves when the number of local updates increases. *That is, the proposed method is able to further reduce the communication cost by reducing the number of rounds required for communication.* This is also consistent with our theoretical findings.

In general, FS−HEAPRIX outperforms all competing methods, and a sketch size of $(50, 100)$ is sufficient to approach the accuracy of full-precision FedSGD.

Heterogeneous case. We plot similar set of results in Figure 2 for non-i.i.d. data distribution, which leads to more twists and turns in the training curves. We see that SketchSGD performs very poorly in the heterogeneous case, which is improved by error tracking and momentum in FetchSGD, as expected. However, both of these methods are worse than our proposed FedSketchGATE methods, which can achieve similar generalization accuracy as full-precision FedSGD, even with small sketch size (i.e., 75× compression). Note that, slower convergence and worse generalization of FedSGD in non-i.i.d. data distribution case is also reported in e.g. (McMahan et al., 2017; Chen et al., 2020).

We also notice in Figure 2 the edge of FS−HEAPRIX over FS−PRIVIX in terms of training loss and test accuracy. However, we see that in the heterogeneous setting, more local updates tend to undermine the learning performance, especially with small sketch size. Nevertheless, when the sketch size is not too small, i.e., $(50, 100)$, FS−HEAPRIX can still provide comparable test accuracy as FedSGD in both cases. Our empirical study demonstrates that FedSketch (and FedSketchGATE) frameworks are able to perform well in homogeneous (resp. heterogeneous) settings, with high compression rate. In particular, FedSketch methods are beneficial over SketchedSGD (Ivkin et al., 2019) and FetchSGD (Rothchild et al., 2020) in all cases. FS−HEAPRIX performs the best among all the tested compressed algorithms, which in many cases achieves similar generalization accuracy as full-precision FedSGD with small sketch size.

6. Conclusion

In this paper, we introduced FedSKETCH and FedSKETCHGATE algorithms for homogeneous and heterogeneous data distribution setting respectively for Federated Learning wherein communication between server and devices is only performed using count sketch. Our algorithms, thus, provide communication-efficiency and privacy, through random hashes based sketches. We analyze the convergence error for *non-convex*, *PL* and *general convex* objective functions in the scope of Federated Optimization. We provide insightful numerical experiments showcasing the advantages of our FedSKETCH and FedSKETCHGATE methods over current federated optimization algorithm. The proposed algorithms outperform competing compression method and can achieve comparable test accuracy as Federated SGD, with high compression ratio.

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 $\{h_{j,1\leq j\leq t} : [d] \rightarrow m\}$ are used along with another set of pairwise independent sign hash functions $\{\text{sign}_{j,1\leq j\leq t} : [d] \rightarrow \{+1, -1\}\}$ 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 $\mathbf{S}(\boldsymbol{x})$:
for $i = 1, \dots, d$ 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 & 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 & Richtárik, 2020) and Algorithm 2 by setting $\Delta_1 = c\frac{d}{m}$ and $\Delta_2 = 1 + c\frac{d}{m}$ we obtain $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} = c\frac{d}{m} = O\left(c\frac{d}{m}\right)$ 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 $\eta = \frac{1}{2L(cd/m+1/k)\tau\gamma}$: $R = O\left(L\left(1/k + cd/m\right)/\epsilon \log\left(1/\epsilon\right)\right)$ and $\tau = O\left(1/\epsilon^2\right)$.*
- *FS-HEAPRIX, for $\eta = \frac{1}{2L((cd-m)/mk+1)\tau\gamma}$: $R = O\left(L\left(1/k + (cd-m)/m\right)/\epsilon \log\left(1/\epsilon\right)\right)$ 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 $\eta = 1/(2L(cd/m+1)\tau\gamma)$: $R = O\left(L(cd/m+1)\epsilon \log(1/\epsilon)\right)$ and $\tau = O\left(1/(p\epsilon^2)\right)$.*
- *FS-HEAPRIX, for $\eta = m/(2Lcd\tau\gamma)$: $R = O\left(Lc(d/m)\epsilon \log(1/\epsilon)\right)$ 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	$\boldsymbol{R = O\left(\frac{1}{\epsilon}\right)}$ $\boldsymbol{\tau = O\left((\mu^2(cd-m) + \frac{\mu^2}{k})\frac{1}{\epsilon}\right)}$ $\boldsymbol{B = O(m \log(\frac{dR}{\delta}))}$ $\boldsymbol{kBR = O(mk/\epsilon \log(\frac{d}{\epsilon\delta}))}$	✔	✗

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 $\gamma = m/d$)	$R = O\left(\frac{d}{m\epsilon^{1.5}}\right)$ $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$ $B = O\left(d\right)$ $RB = O\left(\frac{d^2}{m\epsilon^{1.5}}\right)$	$-$	✗	✗
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	$\boldsymbol{R = O\left(c\frac{d}{m\epsilon}\right)}$ $\boldsymbol{\tau = O(\frac{1}{p\epsilon})}$ $\boldsymbol{B = O(m\log(\frac{cd^2}{m\epsilon\delta}))}$ $\boldsymbol{RB = O(\frac{d}{\epsilon}\log(\frac{cd^2}{m\epsilon\delta}\log(\frac{1}{\epsilon})))}$	$\boldsymbol{R{=}O(\frac{cd}{m\epsilon}\log(\frac{1}{\epsilon}))}$ $\boldsymbol{\tau{=}O(\frac{1}{p\epsilon^2})}$ $\boldsymbol{B = O(m\log(\frac{cd^2}{m\epsilon\delta}))}$	✔	✔

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 $\gamma = \frac{m}{cd}$ (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 $R = O\left(\frac{c\kappa d}{m}\log(1/\epsilon)\right)$ 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 $R = O\left(c\frac{d}{m\epsilon}\right)$ 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 $RB = O\left(\kappa d\log(\frac{c\kappa d^2}{m\delta}\log(\frac{1}{\epsilon}))\log(1/\epsilon)\right) = O\left(\kappa d\max\left(\log(\frac{c\kappa d^2}{m\delta}),\log^2(1/\epsilon)\right)\right)$ and $RB = O\left(\log(c\frac{d^2}{m\epsilon\delta})\frac{d}{\epsilon}\right)$.

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 & Dieuleveut, 2020). The reference (Philippenko & 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 $\omega^{\mathrm{Up}} = \omega^{\mathrm{Down}} = \theta\left(d\right)$) 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 (**?**) and (**?**) that analyze the two-way compression, but since (Philippenko & 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., 2020d; Haddadpour & Mahdavi, 2019)) in proving results.

Fact 5 ((Li et al., 2020d; Haddadpour & 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}}_{\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}\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[\omega\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]\\
&=\omega\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\omega\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]\\
&=\omega\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\omega\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]\\
&=(k\omega+1)\frac{\tau\sigma^2}{k}+(\omega+1)\left[\sum_{j=1}^{p}q_j\|\mathbf{g}_j^{(r)}\|^2\right]
\end{aligned}
\tag{4}
$$

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 $\mathbb{E}_{\mathbf{S}}\left[\frac{1}{p}\sum_{j=1}^{p}\tilde{\mathbf{g}}_{\mathbf{S}j}^{(r)}\right]=\frac{1}{p}\sum_{j=1}^{m}\tilde{\mathbf{g}}_j^{(r)}$.

Next we show that from Assumptions 3, we have

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

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 2.

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

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

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

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

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:

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

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

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

where ① comes from $\mathbb{E}\left[\mathbf{x}^2\right] = \text{Var}\left[\mathbf{x}\right] + \left[\mathbb{E}\left[\mathbf{x}\right]\right]^2$ and ② holds because $\text{Var}\left(\sum_{j=1}^{n}\mathbf{x}_j\right) = \sum_{j=1}^{n}\text{Var}\left(\mathbf{x}_j\right)$ 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*

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

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:*

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

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

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

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:

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

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

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

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

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

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 $\eta\gamma = O\left(\frac{1}{\sqrt{R\tau\left(\omega+\frac{1}{k}\right)}}\right)$.

□

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 & 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 & 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:*

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

Proof. From (18) under condition:

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

we obtain:

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

which leads to the following bound:

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

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

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

□

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:*

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

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 $\tilde{f}(\boldsymbol{w}, \phi) = f(\boldsymbol{w}) + \frac{\phi}{2}\left\|\boldsymbol{w}\right\|^2$, 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:

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

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:

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

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

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

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.

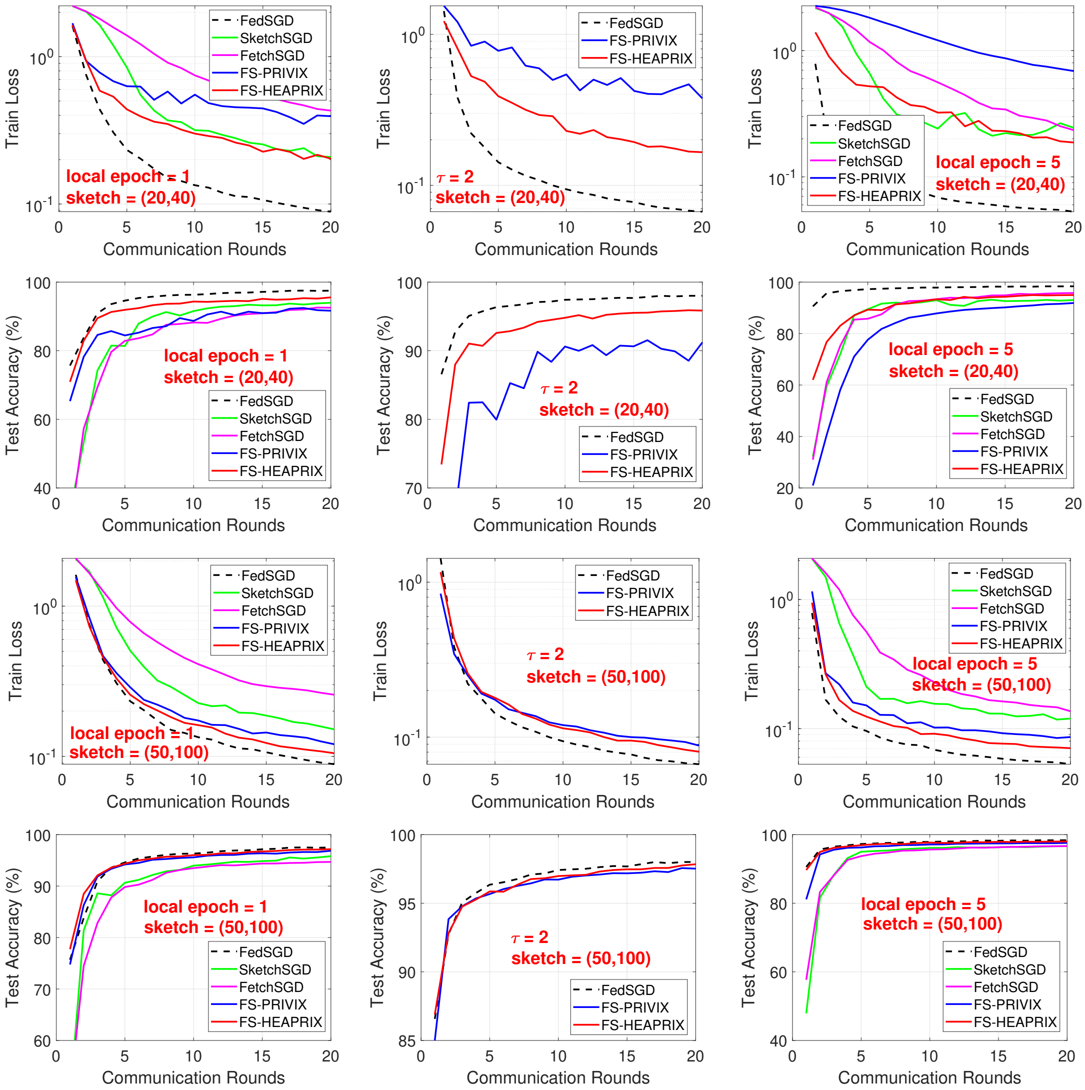

Figure 3. MNIST Homogeneous case: Comparison of compressed optimization methods on LeNet CNN architecture.

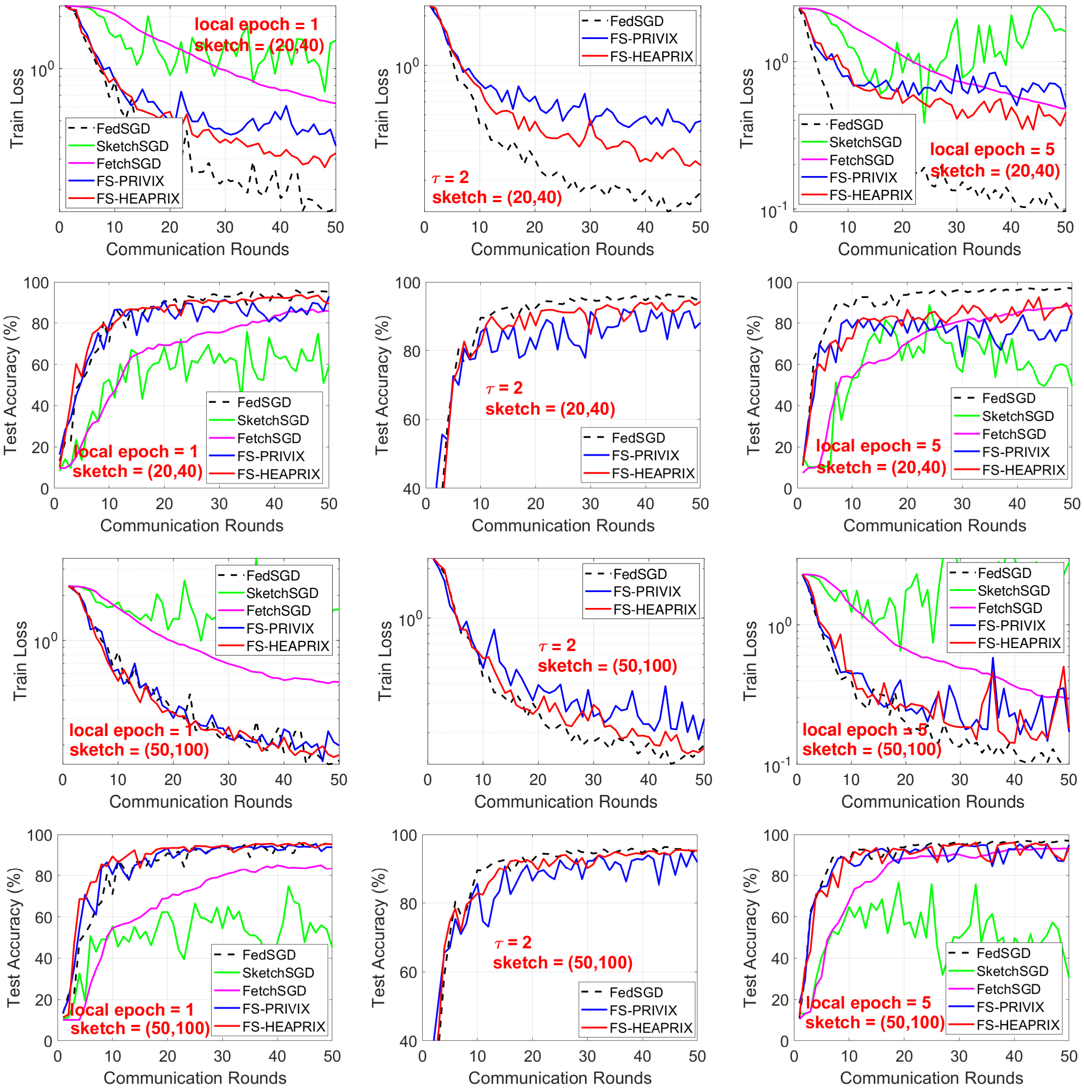

Figure 4. MNIST Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN architecture.

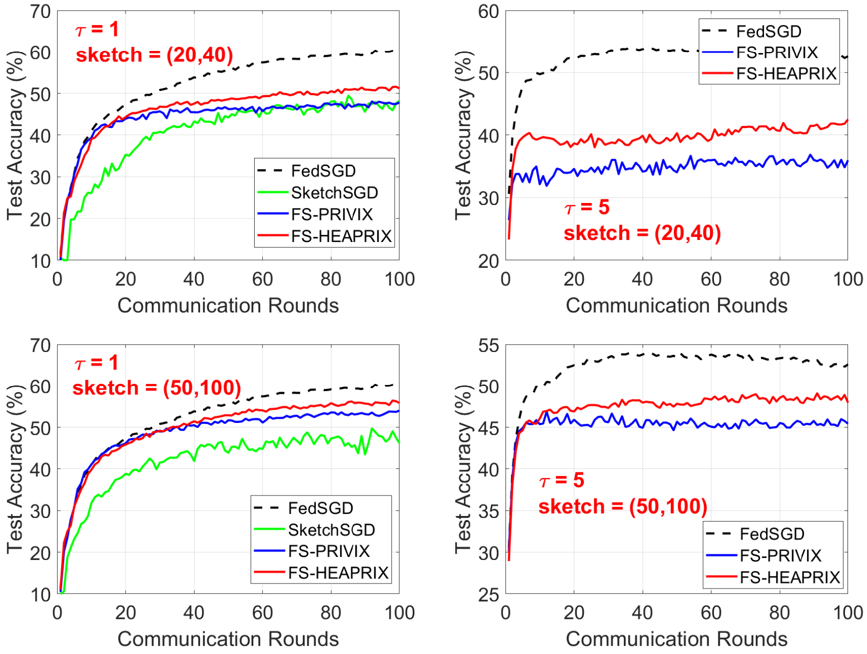

Figure 5. Homogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.

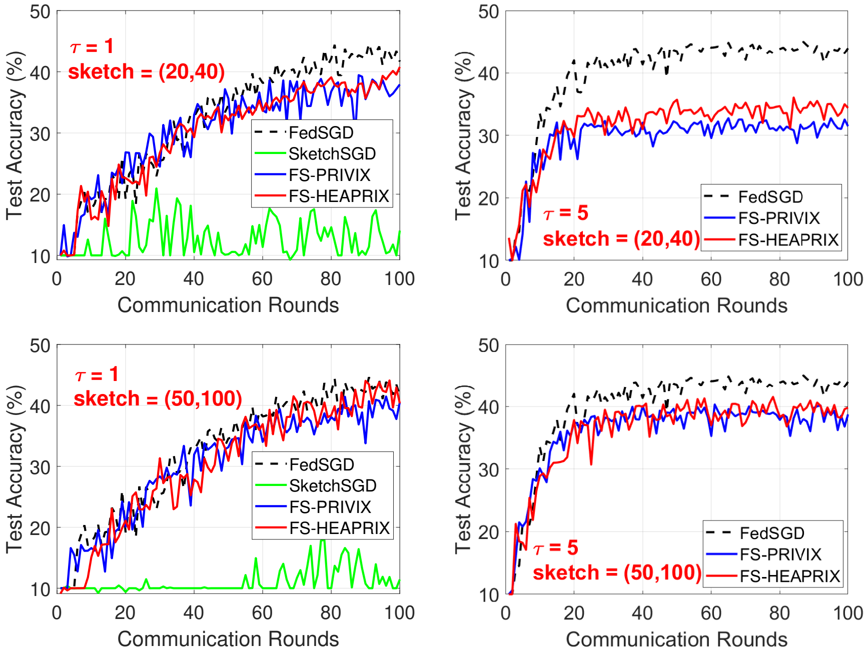

Figure 6. Heterogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.