FedSketch: Communication-Efficient and Private Federated Learning via Sketching

Abstract

Communication complexity and privacy are the two key challenges in Federated Learning where the goal is to perform a distributed learning through a million devices. In this work, we introduce FedSKETCH and FedSKETCHGATE algorithms to address both challenges in Federated learning jointly, where these algorithms are intended to be used for homogenous and heterogenous data distribution settings respectively. The key idea is to compress the accumulation of local gradients using count sketch, therefore, the server does not have access to the gradients themselves which provides privacy. Furthermore, due to the lower dimension of sketching used, we have communication-efficiency property as well. Furthermore, for the aforementioned schemes, we provide sharp convergence guarantees. Finally, we back up our theory with various set of experiments.

1 Introduction

Increasing applications in machine learning include the learning of a complex model across a large amount of devices in a distributed manner. Two natural problems arise from this setting. The first one

The main contributions of this paper are summarized as follows:

- We develop a general algorithm for communication-efficient and privacy preserving federated learning based on a novel compression algorithm. This latter leverage two commonly used compression methods and display an unbiased compressed estimator of the full gradient.
- Based on the current compression methods, we provide

The remaining of the paper is organized as follows. Section **??** gives a formal presentation of the general problem. Section **??** describes the various compression algorithms used for communication efficiency and privacy preservation, and introduces our new compression method. The training algorithms are provided in Section **??** and their respective analysis in the strongly-convex or nonconvex cases are provided Section **??**.

Related Work for Distributed Setting:

Related Work for Privacy-preserving Setting:

2 Problem Setting

In this paper our goal is to solve the following optimization problem using p distributed devices:

$$f(\boldsymbol{x}) \triangleq \left[\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{p} \sum_{j=1}^{p} F_j(\boldsymbol{x}) \right] \tag{1}$$

where $F_j(\boldsymbol{x}) = \mathbb{E}_{\xi\in\mathcal{D}_j}\left[f_j\left(\boldsymbol{x},\xi\right)\right]$ is the local cost function at device j. ξ is a random variable with probability distribution $\mathcal{D}_j$.

ToDo: Differences and potential improvements over [1]!

3 Count Sketch Review

Algorithm 1 `CS`: Count Sketch to compress $\boldsymbol{x} \in \mathbb{R}^d$.

Inputs: $\boldsymbol{x} \in \mathbb{R}^d, t, k, \mathbf{S}_{t\times k}, h_i(1 \le i \le t), sign_i(1 \le i \le t)$
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 $\mathbf{S}_{t\times k}(\boldsymbol{x})$

Notation: For the rest of the paper we indicate the number of communication rounds and number of bits per round per device with $R(\epsilon)$ and $B(d)$ respectively. For the rest of the paper we indicate the count sketch of any vector $\boldsymbol{x}$ with $\mathbf{S}(\boldsymbol{x})$

4 Compression Operations

In this subsection, we review a recent results that will be useful for our work. Similar to [2], we define the following two types of compressor operators that will be useful for our algorithm.

4.1 Unbiased Compressor

Definition 1 (Unbiased compressor)**.** *A randomized function, $C: \mathbb{R}^d \to \mathbb{R}^d$ is called an unbiased compression operator with $\Delta \ge 1$, if we have*

$$\begin{aligned} \mathbb{E}\left[C(\boldsymbol{x})\right] &= \boldsymbol{x} \\ \mathbb{E}\left[\left\|C(\boldsymbol{x})\right\|_2^2\right] &\le \Delta \left\|\boldsymbol{x}\right\|_2^2 \end{aligned} \tag{2}$$

We indicate this class of compressor with $C \in \mathbb{U}(\Delta)$

We note that this definition leads to the property

$$\mathbb{E}\left[\left\|\mathrm{C}(\boldsymbol{x}) - \boldsymbol{x}\right\|_2^2\right] \le (\Delta - 1)\left\|\boldsymbol{x}\right\|_2^2 \tag{3}$$

Remark 1. *Note that in case of $\Delta = 1$ our algorithm reduces for the case of no compression. This property allows us the noise of the compression.*

Algorithm 2 `PRIVIX`[3]: Unbiased compressor based on sketching.

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

Estimation errors:

Property 1 ([3])**.** *For our proof purpose we will need the following crucial properties of the count sketch described in Algorithm 2, for any real valued vector $\mathbf{x} \in \mathbb{R}^d$:*

1) Unbiased estimation*: As it is also mentioned in [3], we have:*

$$\mathbb{E}_{\mathbf{S}}\left[\mathit{PRIVIX}\left[\mathbf{S}\left(\mathbf{x}\right)\right]\right]=\mathbf{x} \tag{4}$$

2) Bounded variance*: With* $k=O\left(\frac{e}{\mu^2}\right)$ *and* $t=O\left(\ln\left(\frac{1}{\delta}\right)\right)$*, we have the following bound with probability* $1-\delta$*:*

$$\mathbb{E}_{\mathbf{S}}\left[\left\|\mathit{PRIVIX}\left[\mathbf{S}\left(\mathbf{x}\right)\right]-\mathbf{x}\right\|_2^2\right]\leq\mu^2 d\left\|\mathbf{x}\right\|_2^2 \tag{5}$$

Therefore, PRIVIX $\in\mathbb{U}(1+\mu^2 d)$ with probability $1-\delta$.

Remark 2. *We note that* $\Delta=1+\mu^2 d$ *implies that if* $k\to d$*,* $\Delta\to 1+1=2$*, which means that the case of no compression is not covered. Thus, the algorithms based on this may converges poorly.*

Differentially Private Property:

Definition 2. *A randomized mechanism* $\mathcal{O}$ *satisfies* $\epsilon-$*differential privacy, if for input data* S_1 *and* S_2 *differing by up to one element, and for any output* D *of* $\mathcal{O}$*,*

$$\Pr\left[\mathcal{O}(S_1)\in D\right]\leq\exp\left(\epsilon\right)\Pr\left[\mathcal{O}(S_2)\in D\right] \tag{6}$$

ToDo: Add explanations that this scheme induces local privacy!

Assumption 1 (Input vector distribution)**.** *For the purpose of privacy analysis, similar to [?, ?], we suppose that for any input vector* S *with length* $|S|=l$*, each element* $s_i\in S$ *is drawn i.i.d. from a Gaussian distribution:* $s_i\sim\mathcal{N}(0,\sigma^2)$*, and bounded by a large probability:* $|s_i|\leq C, 1\leq i\leq p$ *for some positive constant* $C>0$*.*

Theorem 1 ($\epsilon-$ differential privacy of count sketch, [3])**.** *For a sketching algorithm* $\mathcal{O}$ *using Count Sketch* $\mathbf{S}_{t\times k}$ *with* t *arrays of* k *bins, for any input vector* S *with length* l *satisfying Assumption 1,* $\mathcal{O}$ *achieves* $t.\ln\left(1+\frac{\alpha C^2 k(k-1)}{\sigma^2(l-2)}(1+\ln(l-k))\right)-$*differential privacy with high probability, where* α *is a positive constant satisfying* $\frac{\alpha C^2 k(k-1)}{\sigma^2(l-2)}(1+\ln(l-k))\leq\frac{1}{2}-\frac{1}{\alpha}$*.*

The proof of this theorem can be found in [3].

4.2 Biased compressor

Definition 3 (Biased compressor)**.** *A (randomized) function,* $C:\mathbb{R}^d\to\mathbb{R}^d$ *is called a compression operator with* $\alpha>0$ *and* $\Delta\geq 1$*, if we have*

$$\mathbb{E}\left[\left\|\alpha\boldsymbol{x}-\bar{C}(\boldsymbol{x})\right\|_2^2\right]\leq\left(1-\frac{1}{\Delta}\right)\left\|\boldsymbol{x}\right\|_2^2 \tag{7}$$

Any biased compression operator C *is indicated by* $C\in\mathbb{C}(\Delta,\alpha)$*.*

The following Lemma links these two definitions:

Lemma 1 ([2])**.** *We have* $\mathbb{U}(\Delta)\subset\mathbb{C}(\Delta)$*.*

An instance of biased compressor based on sketching is as follows:

Algorithm 3 HEAVYMIX [4]

Inputs: $\mathbf{S}_{\mathbf{g}}$; parameter-k
Compress vector $\tilde{\mathbf{g}}\in\mathbb{R}^d$ **into** $\mathbf{S}\left(\tilde{\mathbf{g}}\right)$**:**
Query $\hat{\ell}_2^2=(1\pm 0.5)\left\|\mathbf{g}\right\|^2$ from sketch $\mathbf{S}_{\mathbf{g}}$
$\forall j$ query $\hat{\mathbf{g}}_j^2=\hat{\mathbf{g}}_j^2\pm\frac{1}{2k}\left\|\mathbf{g}\right\|^2$ from sketch $\mathbf{S}_{\mathbf{g}}$
$H=\{j|\hat{\mathbf{g}}_j\geq\frac{\hat{\ell}_2^2}{k}\}$ and $NH=\{j|\hat{\mathbf{g}}_j<\frac{\hat{\ell}_2^2}{k}\}$
$\mathrm{Top}_k=H\cup rand_\ell(NH)$, where $\ell=k-|H|$
Get exact values of Top_k
Output: $\mathbf{g}_S:\forall j\in\mathrm{Top}_k:\mathbf{g}_{Si}=\mathbf{g}_i$ and $\forall\notin\mathrm{Top}_k:\mathbf{g}_{Si}=0$

ToDo: Explain minor distinction here!

Lemma 2 ([4]). *HEAVYMIX, with sketch size* $\Theta\left(k \log\left(\frac{d}{\delta}\right)\right)$ *is a biased compressor with* $\alpha = 1$ *and* $\Delta = d/k$ *with probability* $\geq 1 - \delta$. *In other words, with probability* $1 - \delta$, *HEAVYMIX* $\in C(\frac{d}{k}, 1)$.

4.3 Sketching Based on Induced Compressor

The following Lemma from [2] shows that how we can transfer biased compressor into an unbiased compressor:

Lemma 3 (Induced Compressor [2]). *For* $C_1 \in \mathbb{C}(\Delta_1)$ *with* $\alpha = 1$, *choose* $C_2 \in \mathbb{U}(\Delta_2)$ *and define the induced compressor with*

$$C(\mathbf{x}) = C_1(\mathbf{x}) + C_2\left(x - C_1\left(\mathbf{x}\right)\right) \tag{8}$$

The induced compressor C *satisfies* $C \in \mathbb{U}(\mathbf{x})$ *with* $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1}$.

Remark 3. *We note that if* $\Delta_2 \geq 1$ *and* $\Delta_1 \leq 1$, *we have* $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1} \leq \Delta_2$

Using this concept of the induced compressor we introduce the following:

Corollary 1. *Based on Lemma 3 and defining*

$$\textit{HEAPRIX}\left[\mathbf{S}\left(\boldsymbol{x}\right)\right] = \textit{HEAVYMIX}\left[\mathbf{S}\left(\mathbf{x}\right)\right] + \textit{PRIVIX}\left[\mathbf{S}\left(\mathbf{x} - \textit{HEAVYMIX}\left[\mathbf{S}\left(\mathbf{x}\right)\right]\right)\right] \tag{9}$$

we have $C(x) \in \mathbb{U}(\mu^2 d)$.

Remark 4. *We highlight that in this case if* $k \to d$, *then* $C(x) \to x$ *which means that your convergence algorithm can be improved by decreasing the noise of compression (with choice of bigger* k*).*

In the following we define two general framework for different sketching algorithms for homogeneous and heterogeneous data distributions.

5 Algorithms for homogeneous and heterogeneous settings

In the following, first we present two algorithm for homogeneous setting. Then, we present two algorithms for heterogeneous algorithms to deal with data heterogeneity.

5.1 Homogeneous setting

In this section, we propose two algorithms for the setting where data at distributed devices is correlated. The proposed Federated Learning with averaging uses sketching to compress communication. The main difference between first algorithm and the algorithm in [3] is that we use distinct local and global learning rates. Additionally, unlike [3] we do not add add local Gaussian noise for the privacy purpose.

In `FedSKETCH-I`, we indicate the number of communication rounds between devices and server with R, and the number of local updates at device j is illustrated with τ, which happens between two consecutive communication rounds. Unlike [1], server node does not store any global model, instead device j has two models, $\boldsymbol{x}^{(r)}$ and $\boldsymbol{x}_j^{(\ell,r)}$. In communication round r device j, the local model $\boldsymbol{x}_j^{(\ell,r)}$ is updated using the rule

$$\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta \tilde{\mathbf{g}}_j^{(\ell,r)}, \qquad \text{for } \ell = 0, \ldots, \tau - 1$$

where $\tilde{\mathbf{g}}_j^{(\ell,r)} \triangleq \nabla f_j(\boldsymbol{x}_j^{(\ell,r)}, \Xi_j^{(\ell,r)}) \triangleq \frac{1}{b} \sum_{\xi \in \Xi_j^{(\ell,r)}} \nabla L_j(\boldsymbol{x}_j^{(\ell,r)}, \xi)$ is a stochastic gradient of f_j evaluated using the mini-batch $\Xi_j^{(\ell,r)} = \{\xi_{j,1}^{(\ell,r)}, \ldots, \xi_{j,b_j}^{(\ell,r)}\}$ of size b_j. η is the local learning rate. After τ local updates locally, model at device j and communication round r is indicated by $\boldsymbol{x}_j^{(\tau,r)}$. The next step of our algorithm is that device j sends the count sketch $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j\left(\boldsymbol{x}_j^{(\tau,r)} - \boldsymbol{x}_j^{(0,r)}\right)$ back to the server. We highlight that

$$\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j\left(\boldsymbol{x}_j^{(\tau,r)} - \boldsymbol{x}_j^{(0,r)}\right) = \mathbf{S}_j\left(\eta \sum_{\ell=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(\ell,r)}\right) = \eta \mathbf{S}_j\left(\sum_{\ell=0}^{\tau-1} \tilde{\mathbf{g}}_j^{(\ell,r)}\right),$$

which is the aggregation of the consecutive stochastic gradients multiplied with local updates η.

Upon receiving all $\mathbf{S}_j^{(r)}$ from devices, the server computes

$$\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1}^{p} \mathbf{S}_j^{(r)} \tag{10}$$

and broadcasts it to all devices. Devices after receiving $\mathbf{S}^{(r)}$ from server updates global model $\boldsymbol{x}^{(r)}$ using rule

$$\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma \texttt{PRIVIX}\left[\mathbf{S}^{(r-1)}\right]$$

All these steps are summarized in `FedSKETCH-I` (Algorithm 4).

Algorithm 4 `FedSKETCH-I`(R, τ, η, γ): Private Federated Learning with Sketching.

Inputs: $\boldsymbol{x}^{(0)}$ as an initial model shared by all local devices, the number of communication rounds R, the the number of local updates τ, and global and local learning rates γ and η, respectively
for $r = 0, \ldots, R-1$ **do**
  parallel for device $j = 1, \ldots, n$ **do**:
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma \texttt{PRIVIX}\left[\mathbf{S}^{(r-1)}\right]$
    Set $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $c = 0, \ldots, \tau - 1$ **do**
      Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)} \triangleq \nabla f_j(\boldsymbol{x}_j^{(\ell,r)}, \xi_j^{(c,r)})$
      $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta\, \tilde{\mathbf{g}}_j^{(\ell,r)}$
    end for
    Device j sends $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$ back to the server.
  Server **computes**
    $\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1} \mathbf{S}_j^{(r)}$ and **broadcasts** $\mathbf{S}^{(r)}$ to all devices.
  end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$

Next, we describe our algorithm using induced sketching.

Algorithm 5 `FedSKETCH-II`$(R,\ \tau, \eta, \gamma)$: Private Federated Learning with Sketching.

Inputs: $\boldsymbol{x}^{(0)}$ as an initial model shared by all local devices, the number of communication rounds R, the the number of local updates τ, and global and local learning rates γ and η, respectively
for $r = 0, \ldots, R-1$ **do**
  parallel for device $j = 1, \ldots, n$ **do**:
    Computes $\boldsymbol{\Phi}^{(r-1)} \triangleq \texttt{HEAVYMIX}(\mathbf{S}^{(r-1)}) + \texttt{PRIVIX}\left[\mathbf{S}^{(r-1)} - \tilde{\mathbf{S}}^{(r-1)}\right]$
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma \boldsymbol{\Phi}^{(r-1)}$
    Set $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $c = 0, \ldots, \tau - 1$ **do**
      Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)} \triangleq \nabla f_j(\boldsymbol{x}_j^{(\ell,r)}, \xi_j^{(c,r)})$
      $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta\ \tilde{\mathbf{g}}_j^{(c,r)}$
    end for
    Device j sends $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j\left(\boldsymbol{x}_j^{(0,r)} - \ \boldsymbol{x}_j^{(\tau,r)}\right)$ back to the server.
  Server **computes**
    $\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1} \mathbf{S}_j^{(r)}$ and **broadcasts** $\mathbf{S}^{(r)}$ to all devices.
  Second round of communication to obtain $\delta_j^{(r)} := \mathbf{S}_j\left[\texttt{HEAVYMIX}(\mathbf{S}^{(r)})\right]$
  Broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{p}\sum_{j=1}^{p} \delta_j^{(r)}$ to devices
  end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$

ToDo: revise it!

5.2 Heterogeneous setting

Algorithm 6 `FedSKETCHGATE-I`$(R,\ \tau, \eta, \gamma)$: Private Federated Learning with Sketching and gradient tracking.

Inputs: $\boldsymbol{x}^{(0)} = \boldsymbol{x}_j^{(0)}$ as an initial model shared by all local devices, the number of communication rounds R, the the number of local updates τ, and global and local learning rates γ and η, respectively
for $r = 0, \ldots, R-1$ **do**
  parallel for device $j = 1, \ldots, n$ **do**:
    Set $\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau}\left(\texttt{PRIVIX}\left(\mathbf{S}^{(r-1)}\right) - \texttt{PRIVIX}\left(\mathbf{S}_j^{(r-1)}\right)\right)$
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma \texttt{PRIVIX}\left(\mathbf{S}^{(r-1)}\right)$
    Set $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $\ell = 0, \ldots, \tau - 1$ **do**
      Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)} \triangleq \nabla f_j(\boldsymbol{x}_j^{(\ell,r)}, \xi_j^{(\ell,r)})$
      $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta\ \left(\tilde{\mathbf{g}}_j^{(\ell,r)} - \mathbf{c}_j^{(r)}\right)$
    end for
    Device j sends $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}\left(\boldsymbol{x}_j^{(0,r)} - \ \boldsymbol{x}_j^{(\tau,r)}\right)$ back to the server.
  Server **computes**
    $\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1} \mathbf{S}_j^{(r)}$ and **broadcasts** $\mathbf{S}^{(r)}$ to all devices.
  end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$

ToDo: revise it!

Algorithm 7 FedSKETCHGATE-II(R, τ, η, γ): Private Federated Learning with Sketching and gradient tracking.

Inputs: $\boldsymbol{x}^{(0)} = \boldsymbol{x}_j^{(0)}$ as an initial model shared by all local devices, the number of communication rounds R, the the number of local updates τ, and global and local learning rates γ and η, respectively
for $r = 0, \ldots, R-1$ **do**
  parallel for device $j = 1, \ldots, n$ **do**:
    Computes $\boldsymbol{\Phi}(\mathbf{S}^{(r-1)}) \triangleq \texttt{HEAVYMIX}(\mathbf{S}^{(r-1)}) + \texttt{PRIVIX}\left[\mathbf{S}^{(r-1)} - \tilde{\mathbf{S}}^{(r-1)}\right]$
    Set $\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau}\left(\boldsymbol{\Phi}\left(\mathbf{S}^{(r-1)}\right) - \boldsymbol{\Phi}\left(\mathbf{S}_j^{(r-1)}\right)\right)$
    Set $\boldsymbol{x}^{(r)} = \boldsymbol{x}^{(r-1)} - \gamma \boldsymbol{\Phi}\left(\mathbf{S}^{(r-1)}\right)$
    Set $\boldsymbol{x}_j^{(0,r)} = \boldsymbol{x}^{(r)}$
    for $\ell = 0, \ldots, \tau - 1$ **do**
      Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)} \triangleq \nabla f_j(\boldsymbol{x}_j^{(\ell,r)}, \xi_j^{(\ell,r)})$
      $\boldsymbol{x}_j^{(\ell+1,r)} = \boldsymbol{x}_j^{(\ell,r)} - \eta \ \left(\tilde{\mathbf{g}}_j^{(\ell,r)} - \mathbf{c}_j^{(r)}\right)$
    end for
    Device j sends $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}\left(\boldsymbol{x}_j^{(0,r)} - \ \boldsymbol{x}_j^{(\tau,r)}\right)$ back to the server.
    Device j computes

$$\boldsymbol{\Phi}\left(\mathbf{S}_j^{(r)}\right) \triangleq \texttt{HEAVYMIX}[\mathbf{S}_j^{(r)}] + \texttt{PRIVIX}\left[\mathbf{S}_j^{(r)} - \mathbf{S}_j{}^{(r)}\left(\texttt{HEAVYMIX}[\mathbf{S}_j{}^{(r)}]\right)\right]$$

  Server **computes**
    $\mathbf{S}^{(r)} = \frac{1}{p}\sum_{j=1} \mathbf{S}_j^{(r)}$ and **broadcasts** $\mathbf{S}^{(r)}$ to all devices.
  Second round of communication to obtain $\delta_j^{(r)} := \mathbf{S}_j\left(\texttt{HEAVYMIX}[\mathbf{S}^{(r)}]\right)$
  Broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{p}\sum_{j=1}^{p} \delta_j^{(r)}$ to devices
  end parallel for
end
Output: $\boldsymbol{x}^{(R-1)}$

5.3 Our algorithms for different sketching schemes

Privacy-preserving algorithm If we set $\texttt{PRIVIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$,...

Privacy-preserving and Communication-efficient algorithm If we set $\texttt{HEAPRIX}\left(\boldsymbol{x}_j^{(0,r)} - \ \boldsymbol{x}_j^{(\tau,r)}\right)$,...

ToDo: Discuss variations and contrast with [4]!

6 Convergence Analysis

6.1 Assumptions

Assumption 2 (Smoothness and Lower Boundedness)**.** *The local objective function $f_j(\cdot)$ of jth device is differentiable for $j \in [m]$ and L-smooth, i.e., $\|\nabla f_j(\boldsymbol{u}) - \nabla f_j(\mathbf{v})\| \leq L\|\boldsymbol{u} - \mathbf{v}\|$, $\forall\ \boldsymbol{u}, \mathbf{v} \in \mathbb{R}^d$. Moreover, the optimal objective function $f(\cdot)$ is bounded below by $f^* = \min_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$.*

Assumption 3 (Polyak-Łojasiewicz)**.** *A function $f(\boldsymbol{x})$ satisfies the Polyak-Łojasiewicz 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.*

6.2 Convergence of FEDSKETCH for homogeneous setting.

Now we focus on the homogeneous case in which the stochastic local gradient of each worker is an unbiased estimator of the global gradient.

Assumption 4 (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 $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{w}; \Xi_j), \mathbb{E}_{\xi_j}[\tilde{\mathbf{g}}_j] = \nabla f(\boldsymbol{w}) = \mathbf{g}$ with the variance bounded is 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 2. *Suppose that the conditions in Assumptions 2-4 hold. Given $0 < k = O\left(\frac{e}{\mu^2}\right) \leq d$, and Consider* `FedSKETCH` *in Algorithm 5 with sketch size $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$. If the local data distributions of all users are identical (homogeneous setting), then with probability $1 - \delta$ we have*

- ***Nonconvex:***
  1) *For the case of $\Phi_{j,\mathbf{S}} =$ `PRIVIX`$\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, by choosing stepsizes as $\eta = \frac{1}{L\gamma}\sqrt{\frac{p}{R\tau\left(\frac{\mu^2 d}{p}+1\right)}}$ and $\gamma \geq m$, the sequence of iterates satisfies $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ if we set $R = O\left(\frac{1}{\epsilon}\right)$ and $\tau = O\left(\frac{\frac{\mu^2 d}{p}+1}{p\epsilon}\right)$.*
  2) *For the case of $\Phi_{j,\mathbf{S}} =$ `HEAPRIX`$\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, by choosing stepsizes as $\eta = \frac{1}{L\gamma}\sqrt{\frac{p}{R\tau\left(\frac{\mu^2 d-1}{p}+1\right)}}$ and $\gamma \geq m$, the sequence of iterates satisfies $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ if we set $R = O\left(\frac{1}{\epsilon}\right)$ and $\tau = O\left(\frac{\frac{\mu^2 d-1}{p}+1}{p\epsilon}\right)$.*
- ***Strongly convex or PL:***
  1) *For the case of $\Phi_{j,\mathbf{S}} =$ `PRIVIX`$\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, by choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{\mu^2 d}{p}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left(\left(\frac{\mu^2 d}{p}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon}\right)$.*
  2) *For the case of*

$$\Phi_{j,\mathbf{S}} = \texttt{HEAVYMIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right) + \texttt{PRIVIX}\left[\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right) - \texttt{HEAVYMIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)\right], \tag{11}$$

  by choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{\mu^2 d-1}{p}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left(\left(\frac{\mu^2 d-1}{p}+1\right)\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon}\right)$.
- ***Convex:***
  1) *For the case of $\Phi_{j,\mathbf{S}} =$ `PRIVIX`$\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, by choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{\mu^2 d}{p}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left(\frac{L\left(1+\frac{\mu^2 d}{p}\right)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon^2}\right)$.*
  2) *For the case of $\Phi_{j,\mathbf{S}} =$ `HEAPRIX`$\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, by choosing stepsizes as $\eta = \frac{1}{2L\left(\frac{\mu^2 d-1}{p}+1\right)\tau\gamma}$ and $\gamma \geq m$, we obtain that the iterates satisfy $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ if we set $R = O\left(\frac{L\left(\frac{\mu^2 d-1}{p}+1\right)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{m\epsilon^2}\right)$.*

Corollary 2 (Total communication cost). *As a consequence of Remark **??**, the total communication cost per-worker becomes*

$$O(RB) = O\left(Rk\log\left(\frac{dR}{\delta}\right)\right) = O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right) \tag{12}$$

We note that this result in addition to improving over the communication complexity of federated learning of the state-of-the-art from $O\left(\frac{d}{\epsilon}\right)$ *in [5, 6, 7] to* $O\left(\frac{kp}{\epsilon}\log\left(\frac{dp}{\epsilon\delta}\right)\right)$*, it also implies differential privacy. As a result, total communication cost is*

$$BpR = O\left(\frac{kp}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right).$$

We note that the state-of-the-art in [5] the total communication cost is

$$BpR = O\left(pd\left(\frac{1}{\epsilon}\right)\right) = O\left(\frac{pd}{\epsilon}\right) \tag{13}$$

We improve this result, in terms of dependency to d*, to*

$$BpR = O\left(\frac{kp}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right) \tag{14}$$

In comparison to [4], we improve the total communication per worker from $RB = O\left(\frac{k}{\epsilon^2}\log\left(\frac{d}{\epsilon^2\delta}\right)\right)$ *to* $RB = O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)$*.*

Remark 5. *It is worthy to note that most of the available communication-efficient algorithm with quantization or compression only consider communication-efficiency from devices to server. However, Algorithm 5 also improves the communication efficiency from server to devices as well.*

Corollary 3 (Total communication cost for PL or strongly convex)**.** *To achieve the convergence error of* ϵ*, we need to have* $R = O\left(\kappa(\frac{\mu^2 d}{p}+1)\log\frac{1}{\epsilon}\right)$ *and* $\tau = \left(\frac{1}{\epsilon}\right)$*. This leads to the total communication cost per worker of*

$$BR = O\left(k\kappa(\frac{\mu^2 d}{p}+1)\log\left(\frac{\kappa(\frac{\mu^2 d^2}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right) \tag{15}$$

As a consequence, the total communication cost becomes:

$$BpR = O\left(k\kappa(\mu^2 d+p)\log\left(\frac{\kappa(\frac{\mu^2 d^2}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right) \tag{16}$$

We note that the state-of-the-art in [5] the total communication cost is

$$BpR = O\left(\kappa pd\log\left(\frac{1}{\epsilon}\right)\right) = O\left(\kappa pd\log\left(\frac{1}{\epsilon}\right)\right) \tag{17}$$

We improve this result, in terms of dependency to d*, to*

$$BpR = O\left(k\kappa(\mu^2 d+p)\log\left(\frac{\kappa(\frac{\mu^2 d}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right) \tag{18}$$

Improving from pd *to* $p+d$*.*

ToDo: Revise this!

6.3 Convergence of FedSKETCHGATE in data heterogeneous setting.

Assumption 5 (Bounded Local Variance)**.** *For all* $j \in [m]$*, we can sample an independent mini-batch* Ξ_j *of size* $|\xi_j| = b$ *and compute an unbiased stochastic gradient* $\tilde{\mathbf{g}}_j = \nabla f_j(\boldsymbol{w};\Xi_j), \mathbb{E}_\xi[\tilde{\mathbf{g}}_j] = \nabla f_j(\boldsymbol{w}) = \mathbf{g}_j$*. Moreover, the variance of local stochastic gradients is bounded above by a constant* σ^2*, i.e.,* $\mathbb{E}_\Xi\left[\|\tilde{\mathbf{g}}_j - \mathbf{g}_j\|^2\right] \leq \sigma^2$*.*

ToDo: Revise this!

Reference	Objective function			UG	PP
	Nonconvex	PL/Strongly Convex	General Convex		
[4]	–	$R = O\left(\frac{\mu^2 d}{\epsilon}\right)$ $\tau = 1$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pRB = O\left(\frac{p\mu^2 d}{\epsilon} k \log\left(\frac{\mu^2 d^2}{\epsilon\delta}\right)\right)$	–	✗	✗
Theorem 2	$R = O\left(\frac{1}{\epsilon}\right)$ $\tau = O\left(\frac{\frac{\mu^2 d}{p}+1}{p\epsilon}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(\frac{pk}{\epsilon} \log\left(\frac{d}{\epsilon\delta}\right)\right)$	$R = O\left(\kappa\left(\frac{\mu^2 d}{p}+1\right)\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{p\epsilon}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(k\kappa(\mu^2 d + p)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\frac{\mu^2 d}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\right)$	$R = O\left(\frac{1+\frac{\mu^2 d}{p}}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{m\epsilon^2}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(\frac{k}{\epsilon}\kappa(\mu^2 d + p)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\frac{\mu^2 d}{p}+d)\log\frac{1}{\epsilon}}{\epsilon\delta}\right)\right)$	✔	✔
Theorem 2	$R = O\left(\frac{1}{\epsilon}\right)$ $\tau = O\left(\frac{\frac{\mu^2 d-1}{p}+1}{p\epsilon}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(\frac{pk}{\epsilon} \log\left(\frac{d}{\epsilon\delta}\right)\right)$	$R = O\left(\kappa\left(\frac{\mu^2 d-1}{p}+1\right)\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{p\epsilon}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(k\kappa(\mu^2 d - 1 + p)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\frac{\mu^2 d-1}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\right)$	$R = O\left(\frac{1+\frac{\mu^2 d-1}{p}}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau = O\left(\frac{1}{m\epsilon^2}\right)$ $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$ $pBR = O\left(\frac{k}{\epsilon}\kappa(\mu^2 d - 1 + p)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\frac{\mu^2 d-1}{p}+d)\log\frac{1}{\epsilon}}{\epsilon\delta}\right)\right)$	✔	✔

Table 1 Comparison of results with compression and periodic averaging in the homogeneous setting. Here, m is the number of devices, q is compression distortion constant, κ is condition number, ϵ is target accuracy, R is the number of communication rounds, and τ is the number of local updates. ToDo: Continue from here!!!!!!!!

Theorem 3. *Suppose that the conditions in Assumptions 2 and 5 hold. Given* $0 < k = O\left(\frac{e}{\mu^2}\right) \leq d$, *and Consider* `FedSKETCHGATE` *in Algorithm 7 with sketch size* $B = O\left(k \log\left(\frac{dR}{\delta}\right)\right)$. *If the local data distributions of all users are identical (homogeneous setting), then with probability* $1 - \delta$ *we have*

- ***Nonconvex:***
  1) *For the case of* $\Phi_{j,\mathbf{S}} = \texttt{PRIVIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, *by choosing stepsizes as* $\eta = \frac{1}{L\gamma}\sqrt{\frac{p}{R\tau(\mu^2 d)}}$ *and* $\gamma \geq m$, *the sequence of iterates satisfies* $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ *if we set* $R = O\left(\frac{\mu^2 d+1}{\epsilon}\right)$ *and* $\tau = O\left(\frac{1}{p\epsilon}\right)$.
  2) *For the case of* $\Phi_{j,\mathbf{S}} = \texttt{HEAPRIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, *by choosing stepsizes as* $\eta = \frac{1}{L\gamma}\sqrt{\frac{p}{R\tau(\mu^2 d)}}$ *and* $\gamma \geq m$, *the sequence of iterates satisfies* $\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{w}^{(r)})\right\|_2^2 \leq \epsilon$ *if we set* $R = O\left(\frac{\mu^2 d}{\epsilon}\right)$ *and* $\tau = O\left(\frac{1}{p\epsilon}\right)$.
- ***PL or Strongly convex:***
  1) *For the case of* $\Phi_{j,\mathbf{S}} = \texttt{PRIVIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, *by choosing stepsizes as* $\eta = \frac{1}{2L(\mu^2 d+1)\tau\gamma}$ *and* $\gamma \geq m$, *we obtain that the iterates satisfy* $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ *if we set* $R = O\left(\left(\mu^2 d + 1\right)\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = O\left(\frac{1}{m\epsilon}\right)$.
  2) *For the case of*

$$\Phi_{j,\mathbf{S}} = \texttt{HEAVYMIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right) + \texttt{PRIVIX}\left[\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right) - \texttt{HEAVYMIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)\right], \tag{19}$$

  by choosing stepsizes as $\eta = \frac{1}{2L(\mu^2 d)\tau\gamma}$ *and* $\gamma \geq m$, *we obtain that the iterates satisfy* $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ *if we set* $R = O\left(\left(\mu^2 d\right)\kappa \log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = O\left(\frac{1}{m\epsilon}\right)$.
- ***Convex:***
  1) *For the case of* $\Phi_{j,\mathbf{S}} = \texttt{PRIVIX}\left(\boldsymbol{x}_j^{(0,r)} - \boldsymbol{x}_j^{(\tau,r)}\right)$, *by choosing stepsizes as* $\eta = \frac{1}{2L(\mu^2 d+1)\tau\gamma}$ *and* $\gamma \geq m$, *we obtain that the iterates satisfy* $\mathbb{E}\left[f(\boldsymbol{w}^{(R)}) - f(\boldsymbol{w}^{(*)})\right] \leq \epsilon$ *if we set* $R = O\left(\frac{L\left(1+\mu^2 d\right)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = O\left(\frac{1}{m\epsilon^2}\right)$.

Reference	Objective function			UG	PP
	Nonconvex	PL/Strongly Convex	General Convex		
[3]	–	–	$R=O\left(\frac{\mu^2 d}{\epsilon^2}\right)$ $\tau=1$ $\mathrm{B}=\mathrm{O}\left(k\log\left(\frac{dR}{\delta}\right)\right)$ $pRB=O\left(\frac{p\mu^2 d}{\epsilon^2}k\log\left(\frac{\mu^2 d^2}{\epsilon^2\delta}\right)\right)$	✗	✔
[5]	$R=O\left(\frac{1}{\epsilon}\right)$ $\tau=O\left(\frac{1}{p\epsilon}\right)$ $B=O\left(d\right)$ $pRB=O\left(\frac{pd}{\epsilon}\right)$	$R=O\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ $\tau=O\left(\frac{1}{p\epsilon}\right)$ $B=O\left(d\right)$ $pRB=O\left(p\kappa d\log\left(\frac{1}{\epsilon}\right)\right)$	$R=O\left(\frac{1}{\epsilon}\right)$ $\tau=O\left(\frac{1}{p\epsilon}\right)$ $B=O\left(d\right)$ $pRB=O\left(\frac{pd}{\epsilon}\right)$	✔	✗
Theorem 3	$\boldsymbol{R=O\left(\frac{\mu^2 d+1}{\epsilon}\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{p\epsilon}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(\frac{p(\mu^2 d+1)k}{\epsilon}\log\left(\frac{\mu^2 d^2+d}{\epsilon\delta}\right)\right)}$	$\boldsymbol{R=O\left(\kappa\left(\mu^2 d+1\right)\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{p\epsilon}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(pk\kappa(\mu^2 d+1)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\mu^2 d^2+d)\log\frac{1}{\epsilon}}{\delta}\right)\right)}$	$\boldsymbol{R=O\left(\frac{1+\mu^2 d}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon^2}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(\frac{k}{\epsilon}p\kappa(\mu^2 d+1)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\mu^2 d^2+d)\log\frac{1}{\epsilon}}{\epsilon\delta}\right)\right)}$	✔	✔
Theorem 3	$\boldsymbol{R=O\left(\frac{\mu^2 d}{\epsilon}\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{p\epsilon}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(\frac{p\mu^2 dk}{\epsilon}\log\left(\frac{\mu^2 d^2}{\epsilon\delta}\right)\right)}$	$\boldsymbol{R=O\left(\kappa\left(\mu^2 d\right)\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{p\epsilon}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(k\kappa(p\mu^2 d)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\mu^2 d^2)\log\frac{1}{\epsilon}}{\delta}\right)\right)}$	$\boldsymbol{R=O\left(\frac{\mu^2 d}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)}$ $\boldsymbol{\tau=O\left(\frac{1}{m\epsilon^2}\right)}$ $\boldsymbol{B=O\left(k\log\left(\frac{dR}{\delta}\right)\right)}$ $\boldsymbol{pBR=O\left(\frac{k}{\epsilon}\kappa p(\mu^2 d)\log\frac{1}{\epsilon}\log\left(\frac{\kappa(\mu^2 d^2+d)\log\frac{1}{\epsilon}}{\epsilon\delta}\right)\right)}$	✔	✔

Table 2 Comparison of results with compression and periodic averaging in the heterogeneous setting. Here, m is the number of devices, q is compression distortion constant, κ is condition number, ϵ is target accuracy, R is the number of communication rounds, and τ is the number of local updates. ToDo: Fixing non-convex results!

2) For the case of $\Phi_{j,\mathbf{S}}=\texttt{HEAPRIX}\left(\boldsymbol{x}_j^{(0,r)}-\boldsymbol{x}_j^{(\tau,r)}\right)$, *by choosing stepsizes as* $\eta=\frac{1}{2L(\mu^2 d)\tau\gamma}$ *and* $\gamma\geq m$, *we obtain that the iterates satisfy* $\mathbb{E}\left[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\right]\leq\epsilon$ *if we set* $R=O\left(\frac{L(\mu^2 d)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau=O\left(\frac{1}{m\epsilon^2}\right)$.

Corollary 4 (Total communication cost)**.** *As a consequence of Remark* **??**, *the total communication cost per-worker becomes*

$$O\left(RB\right)=O\left(Rk\log\left(\frac{dR}{\delta}\right)\right)=O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)\tag{20}$$

We note that this result in addition to improving over the communication complexity of federated learning of the state-of-the-art from $O\left(\frac{d}{\epsilon}\right)$ *in [5, 6, 7] to* $O\left(\frac{kp}{\epsilon}\log\left(\frac{dp}{\epsilon\delta}\right)\right)$, *it also implies differential privacy. As a result, total communication cost is*

$$BpR=O\left(\frac{kp}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right).$$

We note that the state-of-the-art in [5] the total communication cost is

$$BpR=O\left(pd\left(\frac{1}{\epsilon}\right)\right)=O\left(\frac{pd}{\epsilon}\right)\tag{21}$$

We improve this result, in terms of dependency to d, *to*

$$BpR=O\left(\frac{kp}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)\tag{22}$$

In comparison to [4], we improve the total communication per worker from $RB=O\left(\frac{k}{\epsilon^2}\log\left(\frac{d}{\epsilon^2\delta}\right)\right)$ *to* $RB=O\left(\frac{k}{\epsilon}\log\left(\frac{d}{\epsilon\delta}\right)\right)$.

Remark 6. *It is worthy to note that most of the available communication-efficient algorithm with quantization or compression only consider communication-efficiency from devices to server. However, Algorithm 5 also improves the communication efficiency from server to devices as well.*

Corollary 5 (Total communication cost for PL or strongly convex)**.** *To achieve the convergence error of* ϵ, *we need to have* $R=O\left(\kappa(\frac{\mu^2 d}{p}+1)\log\frac{1}{\epsilon}\right)$ *and* $\tau=\left(\frac{1}{\epsilon}\right)$. *This leads to the total communication cost per worker of*

$$BR=O\left(k\kappa(\frac{\mu^2 d}{p}+1)\log\left(\frac{\kappa(\frac{\mu^2 d^2}{p}+d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right)\tag{23}$$

As a consequence, the total communication cost becomes:

$$BpR = O\left(k\kappa(\mu^2 d + p)\log\left(\frac{\kappa(\frac{\mu^2 d^2}{p} + d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right) \tag{24}$$

We note that the state-of-the-art in [5] the total communication cost is

$$BpR = O\left(\kappa pd\log\left(\frac{1}{\epsilon}\right)\right) = O\left(\kappa pd\log\left(\frac{1}{\epsilon}\right)\right) \tag{25}$$

We improve this result, in terms of dependency to d, to

$$BpR = O\left(k\kappa(\mu^2 d + p)\log\left(\frac{\kappa(\frac{\mu^2 d}{p} + d)\log\frac{1}{\epsilon}}{\delta}\right)\log\frac{1}{\epsilon}\right) \tag{26}$$

Improving from pd to $p + d$.

6.4 Convergence of FEDSKETCH in data homogeneous setting.

We note that the main issue with Assumption **??** is that since $d \neq 0$, you can not improve the convergence analysis. For this purpose, we propose Algorithm **??**, where the proposed algorithm is not differentially private.

In this case, we use a different assumption as follows:

Remark 7. *Main distinction of Assumption **??** from **??** is that first we do not need unbiased estimation of compression. Additionally, unlike Assumption **??**, if you let $k = d$, we have $\boldsymbol{x} = Comp_{k=d}(\boldsymbol{x})$.*

We note that Algorithm **??** satisfies this Assumption **??** as shown in [4].

Theorem 4 (General non-convex). *Given $0 < k = O\left(\frac{e}{\mu^2}\right) \leq d$ and running Algorithm 5 with sketch of size $c = O\left(k\log\frac{dR}{\delta}\right)$, under Assumptions 2 and **??**, if*

$$L^2\eta^2\tau^2 + mL\tau\eta\left(1 - \frac{k}{d}\right) + 2\gamma L\eta\tau\left(2 - \frac{k}{d}\right) - 1 \leq 0,\ \eta > \frac{1}{mL\tau}, \tag{27}$$

with probability at least $1 - \delta$, we have:

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \leq \frac{2\mathbb{E}\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^{(*)})\right]}{R\tau\gamma\left(\eta - \frac{1}{\tau mL}\right)} + \frac{2\eta^2\gamma L\left(2 - \frac{k}{d}\right)\frac{\sigma^2}{p}}{\left(\eta - \frac{1}{\tau mL}\right)} + \frac{\eta^3 L^2\tau}{\left(\eta - \frac{1}{\tau mL}\right)}\sigma^2 \tag{28}$$

Remark 8 ($k = d$). *ToDo: TBA...*

Corollary 6 (Learning rate range). *Condition in Eq. (**??**) can further simplified as*

$$\frac{1}{mL\tau} < \eta \leq \frac{-\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right) + \sqrt{\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right)^2 + 4}}{2L\tau} \tag{29}$$

We note that m is a hyperparameter that we choose to pick the feasible range for learning rate. Now, if you set $\eta = \frac{1}{\gamma L}\sqrt{\frac{p}{R\tau\left(2 - \frac{k}{d}\right)}}$ which implies the following:

- $\frac{1}{mL\tau} < \frac{1}{\gamma L}\sqrt{\frac{p}{R\tau\left(2 - \frac{k}{d}\right)}} \implies R < \frac{m^2 p\tau}{\gamma^2\left(2 - \frac{k}{d}\right)}$
- $\frac{1}{\gamma L}\sqrt{\frac{p}{R\tau\left(2 - \frac{k}{d}\right)}} \leq \frac{-\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right) + \sqrt{\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right)^2 + 4}}{2L\tau} \implies R \geq \frac{p\tau}{\gamma^2\left(2 - \frac{k}{d}\right)\left(-\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right) + \sqrt{\left(m - \frac{mk}{d} + 4\gamma - \frac{2\gamma k}{d}\right)^2 + 4}\right)^2}$

Therefore, we have the following range for the choice of R:

$$\frac{p\tau}{\gamma^2\left(2-\frac{k}{d}\right)\left(-\left(m-\frac{mk}{d}+4\gamma-\frac{2\gamma k}{d}\right)+\sqrt{\left(m-\frac{mk}{d}+4\gamma-\frac{2\gamma k}{d}\right)^2+4}\right)^2} \leq R < \frac{m^2 p\tau}{\gamma^2\left(2-\frac{k}{d}\right)} \tag{30}$$

Corollary 7. *Based on Corollary **??**, if we choose $\eta = \frac{1}{\gamma}\sqrt{\frac{p}{R\tau\left(2-\frac{k}{d}\right)}} = \frac{n}{mL\tau}$ which also implies $R = \frac{m^2 p\tau}{\gamma^2 n^2\left(2-\frac{k}{d}\right)}$ with $1 < n < m$, then we have:*

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 &\leq \frac{2\mathbb{E}\left[f(\boldsymbol{x}^{(0)})-f(\boldsymbol{x}^{(*)})\right]}{R\tau\gamma\left(\frac{n-1}{m\tau L}\right)} + \frac{2n^2\gamma L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}}{m^2\tau^2L^2\left(\frac{n-1}{m\tau L}\right)} + \frac{n^3L^2\tau}{m^3\tau^3L^3\left(\frac{n-1}{m\tau L}\right)}\sigma^2 \\ &= \frac{2mL\mathbb{E}\left[f(\boldsymbol{x}^{(0)})-f(\boldsymbol{x}^{(*)})\right]}{(n-1)R\gamma} + \frac{2n^2\gamma\left(2-\frac{k}{d}\right)\sigma^2}{m(n-1)p\tau} + \frac{n^3\sigma^2}{m^2(n-1)\tau}\end{aligned} \tag{31}$$

Based on relation $R = \frac{m^2 p\tau}{\gamma^2 n^2\left(2-\frac{k}{d}\right)}$ if we choose $\tau = \frac{\left(2-\frac{k}{d}\right)}{p\epsilon}$ and $m = np$ and $\gamma = m$ we have:

$$R = \frac{1}{n^2\epsilon}$$

and

$$\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \leq \frac{2\epsilon L\mathbb{E}\left[f(\boldsymbol{x}^{(0)})-f(\boldsymbol{x}^{(*)})\right]}{(n-1)} + \frac{2n\epsilon\sigma^2}{p(n-1)} + \frac{n\epsilon\sigma^2}{p(n-1)\left(2-\frac{k}{d}\right)} \tag{32}$$

Theorem 5 (PL/strongly-convex)**.** *Given $0 < k = O\left(\frac{e}{\mu^2}\right) \leq d$ and running Algorithm 5 with sketch of size $c = O\left(k\log\frac{dR}{\delta}\right)$, under Assumptions 2 and **??**, if*

$$L^2\eta^2\tau^2 + mL\tau\eta\left(1-\frac{k}{d}\right) + 2\gamma L\eta\tau\left(2-\frac{k}{d}\right) - 1 \leq 0,\ \eta > \frac{1}{mL\tau}, \tag{33}$$

with probability at least $1-\delta$, Then for the choice of $\eta = \frac{n}{mL\tau}$, for $m > n > 1$, and the choice of $d\left(1-\frac{1}{3n}\right) \leq k \leq d$ with probability $1-\delta$, we obtain:

$$\mathbb{E}\left[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\right] \leq \exp-\left(\frac{\gamma(n-1)R}{m\kappa}\right)\left[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\right] + \frac{\left(\frac{n^3}{2m^2}+\frac{n^2}{p}\gamma L\left(2-\frac{k}{d}\right)\frac{1}{p}\right)}{\mu\tau(n-1)}\sigma^2 \tag{34}$$

7 Experiments

8 Conclusion

References

[1] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," *arXiv preprint arXiv:2007.01154*, 2020.

[2] S. Horváth and P. Richtárik, "A better alternative to error feedback for communication-efficient distributed learning," *arXiv preprint arXiv:2006.11077*, 2020.

[3] T. Li, Z. Liu, V. Sekar, and V. Smith, "Privacy for free: Communication-efficient learning with differential privacy using sketches," *arXiv preprint arXiv:1911.00972*, 2019.

[4] N. Ivkin, D. Rothchild, E. Ullah, I. Stoica, R. Arora *et al.*, "Communication-efficient distributed sgd with sketching," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 144–13 154.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for on-device federated learning," *arXiv preprint arXiv:1910.06378*, 2019.

[6] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[7] X. Liang, S. Shen, J. Liu, Z. Pan, E. Chen, and Y. Cheng, "Variance reduced local sgd with lower communication complexity," *arXiv preprint arXiv:1912.12844*, 2019.

A Appendix

B Proof of main Theorems

The proof of Theorem 2 follows directly from the results in [1]. For the sake of the completeness we review an assumptions from this reference for the quantiziation with their notation.

Assumption 6 ([1]). *The output of the compression operator $Q(\mathbf{x})$ is an unbiased estimator of its input $\mathbf{x}$, and its variance grows with the squared of the squared of ℓ_2-norm of its argument, i.e., $\mathbb{E}\left[Q(\mathbf{x})\right] = \mathbf{x}$ and $\mathbb{E}\left[\left\|Q(\mathbf{x}) - \mathbf{x}\right\|^2\right] \leq q\left\|\mathbf{x}\right\|^2$.*

B.1 Proof of Theorem 2

Based on Assumption 6 we have:

Theorem 6 ([1]). *Consider `FedCOM` in [1]. Suppose that the conditions in Assumptions 2, 4 and 6 hold. If the local data distributions of all users are identical (homogeneous setting), then we have*

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

Proof. Since the sketching `PRIVIX` and `HEAPRIX`, satisfy the Assumption 6 with $q = \mu^2 d$ and $q = \mu^2 d - 1$ respectively with probablity $1-\delta$. Therefore, all the results in Theorem 2, conclude from Theorem 6 with probability $1-\delta$ and plugging $q = \mu^2 d$ and $q = \mu^2 d - 1$ respectively into the corresponding convergence bounds. □

B.2 Proof of Theorem 3

For the heterogeneous setting, the results in [1] requires the following extra assumption that naturally holds for the sketching:

Assumption 7 ([1]). *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 [1] here as follows:

Theorem 7. *Consider `FedCOMGATE` in [1]. If Assumptions 2, 5, 6 and 7 hold, then even for the case the local data distribution of users are different (heterogeneous setting) we have*

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

- ***Convex:*** *By choosing stepsizes as* $\eta = \frac{1}{2L(q+1)\tau\gamma}$ *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* $R = O\left(\frac{L(1+q)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\tau = O\left(\frac{1}{p\epsilon^2}\right)$.

Proof. Since the sketching `PRIVIX` and `HEAPRIX`, satisfy the Assumption 6 with $q = \mu^2 d$ and $q = \mu^2 d - 1$ respectively with probablity $1 - \delta$. Therefore, all the results in Theorem 3, conclude from Theorem 7 with probability $1 - \delta$ and plugging $q = \mu^2 d$ and $q = \mu^2 d - 1$ respectively into the convergence bounds. □

C Convergence result for `FEDSKETCH` without memory

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

$$f(\boldsymbol{x}^{(r+1)}) - f(\boldsymbol{x}^{(r)}) \leq -\gamma\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\rangle + \frac{\gamma^2 L}{2}\|\tilde{\mathbf{g}}^{(r)}\|^2 \tag{35}$$

We define the following:

$$\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)} = \frac{\eta}{p}\sum_{j=1}^{p}\mathbf{S}\left[\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right] \tag{36}$$

Additionally, we define an auxiliary variable as

$$\tilde{\mathbf{g}}^{(r)} = \frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right] \tag{37}$$

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{x}^{(r+1)}) - f(\boldsymbol{x}^{(r)})\Big]\right] &\leq -\gamma\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\right\rangle\right]\right] + \frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\|^2\right] \\
&= -\gamma\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right]\right] + \gamma\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)} - \tilde{\mathbf{g}}_{\mathbf{S}}^{(r)}\right\rangle\right]\right] \\
&\quad + \frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)} - \tilde{\mathbf{g}}^{(r)} + \tilde{\mathbf{g}}^{(r)}\|^2\right] \\
&\overset{(a)}{=} -\gamma\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right]\right] + \gamma\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \mathbf{g}^{(r)} - \mathbf{g}_{\mathbf{S}}^{(r)}\right\rangle\right]\right] \\
&\quad + \frac{\gamma^2 L}{2}\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)} - \tilde{\mathbf{g}}^{(r)} + \tilde{\mathbf{g}}^{(r)}\|^2\right] \\
&\overset{(b)}{\leq} -\gamma\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right]\right] + \frac{\gamma}{2}\left[\frac{1}{mL}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + mL\mathbb{E}_{\mathbf{S}}\left[\left\|\mathbf{g}^{(r)} - \mathbf{g}_{\mathbf{S}}^{(r)}\right\|_2^2\right]\right] \\
&\quad + \gamma^2 L\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left\|\tilde{\mathbf{g}}_{\mathbf{S}}^{(r)} - \tilde{\mathbf{g}}^{(r)}\right\| + \left\|\tilde{\mathbf{g}}^{(r)}\right\|^2\right] \\
&\overset{(c)}{\leq} -\gamma\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right] + \frac{\gamma}{2}\left[\frac{1}{mL}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + mL\left(1 - \frac{k}{d}\right)\left\|\mathbf{g}^{(r)}\right\|_2^2\right] \\
&\quad + \gamma^2 L\mathbb{E}\left[\left(1 - \frac{k}{d}\right)\left\|\tilde{\mathbf{g}}^{(r)}\right\|_2^2 + \left\|\tilde{\mathbf{g}}^{(r)}\right\|_2^2\right] \\
&\overset{(d)}{=} -\gamma\underbrace{\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}), \tilde{\mathbf{g}}^{(r)}\right\rangle\right]}_{\text{(I)}} + \frac{\gamma}{2mL}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + \frac{mL\gamma}{2}\left(1 - \frac{k}{d}\right)\underbrace{\left\|\mathbf{g}^{(r)}\right\|_2^2}_{\text{(II)}} \\
&\quad + \gamma^2 L\left(2 - \frac{k}{d}\right)\underbrace{\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(r)}\right\|_2^2\right]}_{\text{(III)}}
\end{aligned} \tag{38}$$

To bound term (I) in Eq. (38) we use the combination of Lemmas **??** and **??** we obtain:

$$-\gamma\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{x}^{(r)}),\tilde{\mathbf{g}}^{(r)}\right\rangle\right]\leq\frac{\gamma}{2}\eta\frac{1}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left[-\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2-\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2+L^2\eta^2\sum_{\ell=0}^{\tau-1}\left[\tau\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2+\sigma^2\right]\right] \tag{39}$$

Term (II) can be bounded simply as follows:

$$\begin{aligned}\left\|\mathbf{g}^{(r)}\right\|_2^2&=\left\|\frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|_2^2\\&\leq\frac{\tau\eta^2}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2\end{aligned} \tag{40}$$

Next we bound term (III) using the following lemma:

Lemma 4.

$$\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(r)}\right\|_2^2\right]\leq\frac{\eta^2\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2+\frac{\eta^2\tau}{p}\sigma^2 \tag{41}$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(r)}\right\|_2^2\right]&=\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(r)}-\mathbb{E}\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|_2^2\right]+\left\|\mathbb{E}\left[\tilde{\mathbf{g}}^{(r)}\right]\right\|_2^2\\&=\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^{(r)}-\mathbf{g}^{(r)}\right\|_2^2\right]+\left\|\mathbf{g}^{(r)}\right\|_2^2\\&=\mathbb{E}\left[\left\|\frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\tilde{\mathbf{g}}_j^{(c,r)}\right]-\frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|_2^2\right]+\left\|\frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|_2^2\\&=\frac{\eta^2}{p^2}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right\|_2^2\right]+\left\|\frac{\eta}{p}\sum_{j=1}^{p}\left[\sum_{c=0}^{\tau-1}\mathbf{g}_j^{(c,r)}\right]\right\|_2^2\\&\leq\frac{\eta^2}{p^2}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_j^{(c,r)}-\mathbf{g}_j^{(c,r)}\right\|_2^2\right]+\frac{\eta^2\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2\\&\leq\frac{\eta^2}{p^2}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\sigma^2+\frac{\eta^2\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2\\&=\frac{\eta^2\tau}{p}\sum_{j=1}^{p}\sum_{c=0}^{\tau-1}\left\|\mathbf{g}_j^{(c,r)}\right\|_2^2+\frac{\eta^2\tau}{p}\sigma^2\end{aligned} \tag{42}$$

□

Next, we put all the pieces together as follows:

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

$$
\begin{aligned}
&= -\frac{\tau\eta\gamma}{2}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + \frac{\gamma}{2}\eta\frac{1}{p}\sum_{j=1}^{p}\sum_{\ell=0}^{\tau-1}\left[-\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2 + L^2\eta^2\tau^2\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2\right] + \frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2 \\
&\quad + \frac{\gamma}{2mL}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + \frac{mL\gamma}{2}\left(1-\frac{k}{d}\right)\frac{\tau\eta^2}{p}\sum_{j=1}^{p}\sum_{\ell=0}^{\tau-1}\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2 \\
&\quad + \gamma^2L\left(2-\frac{k}{d}\right)\frac{\eta^2\tau}{p}\sum_{j=1}^{p}\sum_{\ell=0}^{\tau-1}\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2 + \gamma^2L\left(2-\frac{k}{d}\right)\frac{\eta^2\tau}{p}\sigma^2 \\
&= -\left(\frac{\tau\eta\gamma}{2}-\frac{\gamma}{2mL}\right)\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \\
&\quad - \left(\frac{\eta\gamma}{2} - \frac{\eta\gamma}{2}\left(L^2\eta^2\tau^2\right) - \frac{mL\eta\gamma}{2}\left(1-\frac{k}{d}\right)\tau\eta - \gamma^2L\eta^2\tau\left(2-\frac{k}{d}\right)\right)\frac{1}{p}\sum_{j=1}^{p}\sum_{\ell=0}^{\tau-1}\left\|\mathbf{g}_j^{(\ell,r)}\right\|_2^2 \\
&\quad + \frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2 + \gamma^2L\left(2-\frac{k}{d}\right)\frac{\eta^2\tau}{p}\sigma^2 \\
&\overset{(a)}{\leq} -\left(\frac{\tau\eta\gamma}{2}-\frac{\gamma}{2mL}\right)\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 + \frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2 + \tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}
\end{aligned}
\tag{43}
$$

where (a) follows from the learning rate choices of

$$
\frac{\eta\gamma}{2} - \frac{\eta\gamma}{2}\left(L^2\eta^2\tau^2\right) - \frac{mL\eta\gamma}{2}\left(1-\frac{k}{d}\right)\tau\eta - \gamma^2L\eta^2\tau\left(2-\frac{k}{d}\right) \geq 0 \tag{44}
$$

which can be simplified further as follows:

$$
1 - L^2\eta^2\tau^2 - mL\tau\eta\left(1-\frac{k}{d}\right) - 2\gamma L\eta\tau\left(2-\frac{k}{d}\right) \geq 0 \tag{45}
$$

Then using Eq. (43) we obtain:

$$
\frac{\tau\gamma}{2}\left(\eta - \frac{1}{\tau mL}\right)\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \leq \mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[f(\boldsymbol{x}^{(r+1)}) - f(\boldsymbol{x}^{(r)})\right]\right] + \tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p} + \frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2 \tag{46}
$$

which leads to the following bound:

$$
\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \leq \frac{2\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[f(\boldsymbol{x}^{(r+1)}) - f(\boldsymbol{x}^{(r)})\right]\right]}{\tau\gamma\left(\eta - \frac{1}{\tau mL}\right)} + \frac{2\eta^2\gamma L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}}{\left(\eta - \frac{1}{\tau mL}\right)} + \frac{\eta^3L^2\tau}{\left(\eta - \frac{1}{\tau mL}\right)}\sigma^2 \tag{47}
$$

Now averaging over r communication rounds we achieve:

$$
\frac{1}{R}\sum_{r=0}^{R-1}\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2 \leq \frac{2\mathbb{E}\left[\mathbb{E}_{\mathbf{S}}\left[f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^{(*)})\right]\right]}{R\tau\gamma\left(\eta - \frac{1}{\tau mL}\right)} + \frac{2\eta^2\gamma L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}}{\left(\eta - \frac{1}{\tau mL}\right)} + \frac{\eta^3L^2\tau}{\left(\eta - \frac{1}{\tau mL}\right)}\sigma^2 \tag{48}
$$

We note that for this case we have the following conditions over learning rate:

$$
L^2\eta^2\tau^2 + mL\tau\eta\left(1-\frac{k}{d}\right) + 2\gamma L\eta\tau\left(2-\frac{k}{d}\right) \leq 1,\ \eta > \frac{1}{mL\tau}, \tag{49}
$$

C.1 Proof of Theorem ??

From Eq. (43) under condition with:

$$
L^2\eta^2\tau^2 + mL\tau\eta\left(1-\frac{k}{d}\right) + 2\gamma L\eta\tau\left(2-\frac{k}{d}\right) \leq 1, \tag{50}
$$

we obtain:

$$\mathbb{E}\left[f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(r)})\right] \leq -\left(\frac{\tau\eta\gamma}{2}-\frac{\gamma}{2mL}\right)\left\|\nabla f(\boldsymbol{x}^{(r)})\right\|_2^2+\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}$$
$$\overset{(PL)}{\leq} -\left(\tau\mu\eta\gamma-\frac{\mu\gamma}{mL}\right)\left[f(\boldsymbol{w}^{(r)})-f(\boldsymbol{w}^{(*)})\right]+\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p} \quad (51)$$

which leads to the following bound:

$$\mathbb{E}\Big[f(\boldsymbol{w}^{(r+1)})-f(\boldsymbol{w}^{(*)})\Big] \leq \left(1-\eta\mu\gamma\tau+\frac{\mu\gamma}{mL}\right)\left[f(\boldsymbol{w}^{(r)})-f(\boldsymbol{w}^{(*)})\right]+\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p} \quad (52)$$

which leads to the following bound by setting $\Delta \triangleq 1-\eta\mu\gamma\tau+\frac{\mu\gamma}{mL}=1-\mu\gamma\tau\left(\eta-\frac{1}{mL\tau}\right)$:

$$\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big] &\leq \Delta^R\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{1-\Delta^R}{1-\Delta}\left(\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}\right)\\
&\leq \Delta^R\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{1}{1-\Delta}\left(\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}\right)\\
&= \left(1-\mu\gamma\tau\left(\eta-\frac{1}{mL\tau}\right)\right)^R\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{\left(\frac{\gamma\eta^3L^2\tau^2}{2}\sigma^2+\tau\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}\right)}{\mu\gamma\tau\left(\eta-\frac{1}{mL\tau}\right)}\\
&\leq \exp-\left(\mu\gamma\tau\left(\eta-\frac{1}{mL\tau}\right)R\right)\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{\left(\frac{\gamma\eta^3L^2\tau}{2}\sigma^2+\eta^2\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}\right)}{\mu\gamma\left(\eta-\frac{1}{mL\tau}\right)}
\end{aligned} \quad (53)$$

Then for the choice of $\eta=\frac{n}{mL\tau}$, for $m>n>1$, we obtain:

$$\begin{aligned}
\mathbb{E}\Big[f(\boldsymbol{w}^{(R)})-f(\boldsymbol{w}^{(*)})\Big] &\leq \exp-\left(\frac{\gamma\left(n-1\right)R}{m\kappa}\right)\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{\left(\frac{\gamma n^3L^2\tau}{2m^3L^3\tau^3}\sigma^2+\frac{n^2}{m^2L^2\tau^2}\gamma^2L\left(2-\frac{k}{d}\right)\frac{\sigma^2}{p}\right)}{\mu\gamma\left(\frac{n-1}{mL\tau}\right)}\\
&= \exp-\left(\frac{\gamma\left(n-1\right)R}{m\kappa}\right)\Big[f(\boldsymbol{w}^{(0)})-f(\boldsymbol{w}^{(*)})\Big]+\frac{\left(\frac{n^3}{2m^2}+\frac{n^2}{m}\gamma L\left(2-\frac{k}{d}\right)\frac{1}{p}\right)}{\mu\tau\left(n-1\right)}\sigma^2
\end{aligned} \quad (54)$$

We note that regarding condition in Eq. (50), if we let $\eta=\frac{n}{mL\tau}$ for $m>n>1$, we need to satisfy the following condition:

$$\frac{n^2}{m^2}+n\left(1-\frac{k}{d}\right)+\frac{2n\gamma\left(1-\frac{k}{d}\right)}{m}\leq 1 \quad (55)$$

Now if you let $\gamma=\frac{m}{2}$, we need to impose the following condition over k and d as follows:

$$n\left(1-\frac{k}{d}\right)\leq\frac{1}{3} \implies d\left(1-\frac{1}{3n}\right)\leq k\leq d \quad (56)$$

ToDo: Will fix these later!