
FedSKETCH: Communication-Efficient Federated Learning via Sketching

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 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
2 large volume of devices. In this work, we introduce two new algorithms, namely
3 FedSKETCH and FedSKETCHGATE, to address jointly both challenges and which
4 are, respectively, intended to be used for homogeneous and heterogeneous data distribution settings. Our algorithms are based on a key and novel sketching technique,
5 called HEAPRIX that is unbiased, compresses the accumulation of local gradients using count sketch, and exhibits communication-efficiency properties leveraging
6 low-dimensional sketches. We provide sharp convergence guarantees of our algorithms and validate our theoretical findings with various sets of experiments.

1 Introduction

12 Federated Learning (FL) is a recently emerging framework for distributed large scale machine learning problems. In FL, data is distributed across devices [33; 23] and due to privacy concerns,
13 users are only allowed to communicate with the parameter server. Formally, the optimization problem across p distributed devices is defined as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^d, \sum_{j=1}^p q_j = 1} f(\mathbf{x}) \triangleq \sum_{j=1}^p q_j F_j(\mathbf{x}), \quad (1)$$

16 where $F_j(\mathbf{x}) = \mathbb{E}_{\xi \in \mathcal{D}_j} [L_j(\mathbf{x}, \xi)]$ is the local cost function at device j , $q_j \triangleq \frac{n_j}{n}$, n_j is the number of data shards at device j and $n = \sum_{j=1}^p n_j$ is the total number of data samples, ξ is a random variable distributed according to probability distribution \mathcal{D}_j , and L_j is a loss function that measures the performance of model \mathbf{x} at device j . We note that, while for the homogeneous setting 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.

22 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:

24 – *Communication-efficiency*: There are often many devices communicating with the server, thus incurring immense communication overhead. One approach to reduce communication round is using
25 local SGD with periodic averaging [48; 41; 47; 43] which periodically averages models after few
26 local updates, contrary to baseline SGD [6] where model averaging is performed at each iteration.
27 Local SGD has been proposed in McMahan et al. [33]; Konečný et al. [23] under the FL setting and its convergence analysis is studied in Stich [41]; Wang and Joshi [43]; Zhou and Cong [48]; Yu et al. [47], later on improved in the follow up references [3; 12; 21; 39] for homogeneous setting. It is
28 further extended to heterogeneous setting [46; 30; 38; 31; 12; 20]. Second approach to deal with
29 communication cost aims at reducing the size of communicated message per communication round,
30 such as local gradient quantization [1; 4; 42; 44; 45] or sparsification [2; 32; 40; 39].

34 *–Data heterogeneity:* Since locally generated data in each device may come from different distribution,
 35 local computations involved in FL setting can lead to poor convergence error in practice [27; 31].
 36 To mitigate the negative impact of data heterogeneity, [13; 16; 31; 20] suggest applying variance
 37 reduction or gradient tracking techniques along local computations.

38 *–Privacy* [11; 14]: Privacy has been widely addressed by injecting an additional layer of randomness
 39 to respect differential-privacy property [34] or using cryptography-based approaches under secure
 40 multi-party computation [5]. Further study of challenges can be found in recent surveys [28] and [18].

41 To tackle all major aforementioned challenges in FL jointly, sketching based algorithms [7; 9; 22; 25]
 42 are promising approaches. For instance, to reduce communication cost, [17] develop a distributed
 43 SGD algorithm using sketching along providing its convergence analysis in the homogeneous setting,
 44 and establish a communication complexity of order $\mathcal{O}(\log(d))$ per round, where d is the dimension
 45 of the vector of parameters compared to $\mathcal{O}(d)$ complexity per round of baseline mini-batch SGD. Yet,
 46 the proposed sketching scheme in Ivkin et al. [17], built from a communication-efficiency perspective,
 47 is based on a deterministic procedure which requires access to the exact information of the gradients,
 48 thus not meeting the crucial privacy-preserving criteria. This systemic flaw is partially addressed
 49 in Rothchild et al. [37].

50 Focusing on privacy, [26] derive a single framework in order to tackle these issues jointly and
 51 introduces DiffSketch algorithm, based on the Count Sketch operator, yet does not provide its
 52 convergence analysis. Additionally, the estimation error of DiffSketch is higher than the sketching
 53 scheme in Ivkin et al. [17] which may end up in poor convergence.

54 In this paper, we propose new sketching algorithms to address the aforementioned challenges
 55 simultaneously. Our main contributions are summarized as:

- 56 • We provide a new algorithm – HEAPRIX – and theoretically show that it reduces the cost
 57 of communication between devices and server, which is based on unbiased sketching with-
 58 out requiring the broadcast of exact values of gradients to the server. Based on HEAPRIX,
 59 we develop general algorithms for communication-efficient and sketch-based FL, namely
 60 FedSKETCH and FedSKETCHGATE for both homogeneous and heterogeneous data distribu-
 61 tion settings respectively.
- 62 • We establish non-asymptotic convergence bounds for convex, Polyak-Łojasiewicz (PL) and
 63 non-convex functions in Theorems 1 and 2 in both homogeneous and heterogeneous cases,
 64 and highlight an improvement in the number of iteration to reach a stationary point. We also
 65 provide a convergence analysis for the PRIVIX algorithm proposed in Li et al. [26].
- 66 • We illustrate the benefits of FedSKETCH and FedSKETCHGATE over baseline methods through
 67 a set of experiments. The latter shows the advantages of the HEAPRIX compression method
 68 achieving comparable test accuracy as Federated SGD (FedSGD) while compressing the
 69 information exchanged between devices and server.

70 **Notation:** We denote the number of communication rounds and bits per round and per device by
 71 R and B respectively. The count sketch of any vector \mathbf{x} is designated by $\mathbf{S}(\mathbf{x})$. $[p]$ denotes the set
 72 $\{1, \dots, p\}$.

73 2 Compression using Count Sketch

74 In this paper, we exploit the commonly used Count Sketch [7] which uses two sets of functions
 75 that encode any input vector \mathbf{x} into a hash table $\mathbf{S}_{m \times t}(\mathbf{x})$. Pairwise independent hash functions
 76 $\{h_{j,1 \leq j \leq t} : [d] \rightarrow [m]\}$ are used along with another set of pairwise independent sign hash functions
 77 $\{\text{sign}_{j,1 \leq j \leq t} : [d] \rightarrow \{+1, -1\}\}$ to map entries of \mathbf{x} ($x_i, 1 \leq i \leq d$) into t different columns of
 78 $\mathbf{S}_{m \times t}$, wherein to lower the dimension of the input vector we usually have $d \gg mt$. The final update
 79 reads $\mathbf{S}[j][h_j(i)] = \mathbf{S}[j-1][h_{j-1}(i)] + \text{sign}_j(i) \cdot x_i$ for any $1 \leq j \leq t$. There are various types of
 80 sketching algorithms which are developed based on count sketching that we develop in the following
 81 subsections. See the Appendix for the detailed Count Sketch algorithm.

82 2.1 Sketching based Unbiased Compressor

83 We define an unbiased compressor as follows:

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

$$\mathbb{E}[C(\mathbf{x})] = \mathbf{x} \quad \text{and} \quad \mathbb{E}[\|C(\mathbf{x})\|_2^2] \leq \Delta \|\mathbf{x}\|_2^2.$$

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

87 This definition leads to the following property

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

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

90 An instance of such unbiased compressor is PRIVIX which obtains an estimate of input \mathbf{x} from a
 91 count sketch noted $\mathbf{S}(\mathbf{x})$. In this algorithm, to query the quantity x_i , the i -th element of the vector
 92 \mathbf{x} , we compute the median of t approximated values specified by the indices of $h_j(i)$ for $1 \leq j \leq t$,
 93 see [26] or Algorithm 6 in the Appendix (for more details). For the purpose of our proof, we state the
 94 following crucial properties of the count sketch:

95 **Property 1** (Li et al. [26]). *For any $\mathbf{x} \in \mathbb{R}^d$, we have:*

96 *Unbiased estimation: As in Li et al. [26], we have $\mathbb{E}_{\mathbf{S}}[\text{PRIVIX}[\mathbf{S}(\mathbf{x})]] = \mathbf{x}$.*

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

$$\mathbb{E}_{\mathbf{S}}[\|\text{PRIVIX}[\mathbf{S}(\mathbf{x})] - \mathbf{x}\|_2^2] \leq c \frac{d}{m} \|\mathbf{x}\|_2^2,$$

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

99 Thus, with probability $1 - \delta$ we obtain that $\text{PRIVIX} \in \mathbb{U}(1 + c \frac{d}{m})$. Note $\Delta = 1 + c \frac{d}{m}$ implies that if
 100 $m \rightarrow d$, then $\Delta \rightarrow 1 + c$, indicating a noisy reconstruction. Exploiting this noisy reconstruction, Li
 101 et al. [26] show that if the data is normally distributed, PRIVIX is differentially private [10], up to
 102 additional assumptions and algorithmic design.

103 2.2 Sketching based Biased Compressor

104 A biased compressor is defined as follows:

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

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

107 The reference [15] proves that $\mathbb{U}(\Delta) \subset \mathbb{C}(\Delta, \alpha)$. An example of bi-
 108 ased compression via sketching and using top_m operation is given below:
 109

110 Following Ivkin et al. [17], HEAVYMIX with
 111 sketch size $\Theta(m \log(\frac{d}{\delta}))$ is a biased compres-
 112 sor with $\alpha = 1$ and $\Delta = d/m$ with probabili-
 113 ty $\geq 1 - \delta$. In other words, with probability
 114 $1 - \delta$, $\text{HEAVYMIX} \in \mathbb{C}(\frac{d}{m}, 1)$. We note that Algo-
 115 rithm 1 is a variation of the sketching algorithm
 116 developed in Ivkin et al. [17] with distinction
 117 that HEAVYMIX does not require a second round
 118 of communication to obtain the exact values of
 119 top_m . Additionally, while a sketching algorithm
 120 implementing HEAVYMIX has smaller estimation
 121 error compared to PRIVIX, it requires having
 122 access to the exact values of top_m , therefore not
 123 benefiting from privacy properties contrary to PRIVIX. In the following we introduce our sketching
 124 scheme – HEAPRIX – as a combination of those two methods.

Algorithm 1 HEAVYMIX

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

2.3 Sketching based Induced Compressor

Due to Theorem 3 in Horváth and Richtárik [15], which illustrates that we can convert the biased compressor into an unbiased one such that, for $C_1 \in \mathbb{C}(\Delta_1)$ with $\alpha = 1$, if you choose $C_2 \in \mathbb{U}(\Delta_2)$, then induced compressor $C : x \mapsto C_1(x) + C_2(x - C_1(x))$ belongs to $\mathbb{U}(\Delta)$ with $\Delta = \Delta_2 + \frac{1-\Delta_2}{\Delta_1}$.

Based on this notion, Algorithm 2 proposes an induced sketching algorithm by utilizing HEAVYMIX and PRIVIX for C_1 and C_2 respectively where the reconstruction of input x is performed using hash table S and x , similar to PRIVIX and HEAVYMIX. Note that if $m \rightarrow d$, then $C(x) \rightarrow x$, which implies that the convergence rate of the algorithm can be improved by decreasing the size of compression m .

Algorithm 2 HEAPRIX

- 1: **Inputs:** $x \in \mathbb{R}^d, t, m, S_{m \times t}, h_j(1 \leq i \leq t), \text{sign}_j(1 \leq i \leq t)$, parameter m
 - 2: Approximate $S(x)$ using HEAVYMIX
 - 3: Approximate $S(x - \text{HEAVYMIX}[S(x)])$ using PRIVIX
 - 4: **Output:**
 $\text{HEAVYMIX}[S(x)] + \text{PRIVIX}[S(x - \text{HEAVYMIX}[S(x)])]$.
-

Corollary 1. *Based on Theorem 3 of [15], HEAPRIX in Algorithm 2 satisfies $C(x) \in \mathbb{U}(c \frac{d}{m})$.*

Benefits of HEAPRIX: Corollary 1 states that, unlike PRIVIX, HEAPRIX compression noise can be made as small as possible using larger hash size. Contrary to HEAVYMIX, HEAPRIX does not require having access to exact top_m values of the input, thus helps preserving privacy. In other words, HEAPRIX leverages the best of both worlds: the *unbiasedness* of PRIVIX while using *heavy hitters* as in HEAVYMIX.

3 FedSKETCH and FedSKETCHGATE

We introduce our 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. [13], server node does not store any global model, rather, device j has two models: $x^{(r)}$ and $x_j^{(\ell, r)}$, which are respectively the local and global models. We develop FedSKETCH in Algorithm 3. A variant of this algorithm implementing HEAPRIX is also described in Algorithm 3. We note that for this variant, we need to have an additional communication round between server and worker j to aggregate $\delta_j^{(r)} \triangleq S_j[\text{HEAVYMIX}(S^{(r)})]$, see Lines 3 and 3. The main difference between our FedSKETCH and the DiffSketch algorithm in Li et al. [26] is that we use distinct local and global learning rates. Furthermore, unlike Li et al. [26], we do not add local Gaussian noise.

Algorithmic comparison with Haddadpour et al. [13] An important feature of our algorithm is that due to a lower dimension of the count sketch, the resulting averages ($S^{(r)}$ and $\tilde{S}^{(r)}$) received by the server, are also of lower dimension. Therefore, these algorithms exploit a bidirectional compression during the communication from server to device back and forth. As a result, due to this bidirectional property of communicating sketching for the case of large quantization error $\omega = \theta(\frac{d}{m})$ as shown in Haddadpour et al. [13], our algorithms can outperform FedCOM and FedCOMGATE developed in Haddadpour et al. [13] if sufficiently large hash tables are used and the uplink communication cost is high. Furthermore, while, in Haddadpour et al. [13], 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. Thus, sketching-based server-devices communication algorithms such as ours do not reveal the exact values of the inputs, to preserve privacy as a by-product.

Remark 1. *As pointed out in Horváth and Richtárik [15], while induced compressors transform a biased compressor into unbiased one, as a drawback it doubles communication cost since the devices need to send $C_1(x)$ and $C_2(x - C_1(x))$ separately. We note that in the special case of HEAPRIX, 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 sketching.*

Algorithm 3 FedSKETCH(R, τ, η, γ)

1: **Inputs:** $\mathbf{x}^{(0)}$: initial model shared by all local devices, global and local learning rates γ and η , respectively
2: **for** $r = 0, \dots, R - 1$ **do**
3: **parallel for device** $j \in \mathcal{K}^{(r)}$ **do**:
4: **if PRIVIX variant:**
$$\Phi^{(r)} \triangleq \text{PRIVIX} \left[\mathbf{S}^{(r-1)} \right]$$

5: **if HEAPRIX variant:**
$$\Phi^{(r)} \triangleq \text{HEAVYMIX} \left[\mathbf{S}^{(r-1)} \right] + \text{PRIVIX} \left[\mathbf{S}^{(r-1)} - \tilde{\mathbf{S}}^{(r-1)} \right]$$

6: Set $\mathbf{x}^{(r)} = \mathbf{x}^{(r-1)} - \gamma \Phi^{(r)}$ and $\mathbf{x}_j^{(0,r)} = \mathbf{x}^{(r)}$
7: **for** $\ell = 0, \dots, \tau - 1$ **do**
8: Sample a mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)}$
9: Update $\mathbf{x}_j^{(\ell+1,r)} = \mathbf{x}_j^{(\ell,r)} - \eta \tilde{\mathbf{g}}_j^{(\ell,r)}$
10: **end for**
11: Device j broadcasts $\mathbf{S}_j^{(r)} \triangleq \mathbf{S}_j \left(\mathbf{x}_j^{(0,r)} - \mathbf{x}_j^{(\tau,r)} \right)$.
12: Server **computes** $\mathbf{S}^{(r)} = \frac{1}{k} \sum_{j \in \mathcal{K}} \mathbf{S}_j^{(r)}$.
13: Server **broadcasts** $\mathbf{S}^{(r)}$ to devices in randomly drawn devices $\mathcal{K}^{(r)}$.
14: **if HEAPRIX variant:**
15: Second round of communication: $\delta_j^{(r)} := \mathbf{S}_j \left[\text{HEAVYMIX}(\mathbf{S}^{(r)}) \right]$ and broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{k} \sum_{j \in \mathcal{K}} \delta_j^{(r)}$ to devices in set $\mathcal{K}^{(r)}$
16: **end parallel for**
17: **end**
18: **Output:** $\mathbf{x}^{(R-1)}$

3.2 Heterogeneous Setting

In this section, we focus on the optimization problem of (1) in the special case of $q_1 = \dots = q_p = \frac{1}{p}$ with full device participation ($k = p$). These results can be extended to the scenario where devices are sampled. 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. Therefore, to mitigate the effect of data heterogeneity, we introduce a new algorithm called FedSKETCHGATE described in Algorithm 4. This algorithm leverages the idea of gradient tracking applied in Haddadpour et al. [13] (with compression) and a special case of $\gamma = 1$ without compression [31]. The main idea is that using an approximation of global gradient, $\mathbf{c}_j^{(r)}$ allows to correct the local gradient direction. For the FedSKETCHGATE with PRIVIX variant, the correction vector $\mathbf{c}_j^{(r)}$ at device j and communication round r is computed in Line 4. While using HEAPRIX compression, FedSKETCHGATE also updates $\tilde{\mathbf{S}}^{(r)}$ via Line 4.

Remark 2. *Most of the existing communication-efficient algorithms with compression only consider communication-efficiency from devices to server. However, Algorithms 3 and 4 also improve the communication efficiency from server to devices since it exploits low-dimensional sketches (and averages), communicated from the server to devices.*

For both FedSKETCH and FedSKETCHGATE algorithms, unlike PRIVIX, HEAPRIX variant requires a second round of communication. Therefore, in Cross-Device FL setting, where there could be millions of devices, HEAPRIX 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).

Algorithm 4 FedSKETCHGATE(R, τ, η, γ)

1: **Inputs:** $\mathbf{x}^{(0)} = \mathbf{x}_j^{(0)}$ shared by all local devices, global and local learning rates γ and η .
 2: **for** $r = 0, \dots, R - 1$ **do**
 3: **parallel for device** $j = 1, \dots, p$ **do**:
 4: **if PRIVIX variant:**

$$\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau} \left[\text{PRIVIX} \left(\mathbf{S}^{(r-1)} \right) - \text{PRIVIX} \left(\mathbf{S}_j^{(r-1)} \right) \right]$$

 5: where $\Phi^{(r)} \triangleq \text{PRIVIX}(\mathbf{S}^{(r-1)})$
 6: **if HEAPRIX variant:**

$$\mathbf{c}_j^{(r)} = \mathbf{c}_j^{(r-1)} - \frac{1}{\tau} \left(\Phi^{(r)} - \Phi_j^{(r)} \right)$$

 7: Set $\mathbf{x}^{(r)} = \mathbf{x}^{(r-1)} - \gamma \Phi^{(r)}$ and $\mathbf{x}_j^{(0,r)} = \mathbf{x}^{(r)}$
 8: **for** $\ell = 0, \dots, \tau - 1$ **do**
 9: Sample mini-batch $\xi_j^{(\ell,r)}$ and compute $\tilde{\mathbf{g}}_j^{(\ell,r)}$
 10: $\mathbf{x}_j^{(\ell+1,r)} = \mathbf{x}_j^{(\ell,r)} - \eta \left(\tilde{\mathbf{g}}_j^{(\ell,r)} - \mathbf{c}_j^{(r)} \right)$
 11: **end for**
 12: Device j broadcasts $\mathbf{S}_j^{(r)} \triangleq \mathbf{S} \left(\mathbf{x}_j^{(0,r)} - \mathbf{x}_j^{(\tau,r)} \right)$.
 13: Server computes $\mathbf{S}^{(r)} = \frac{1}{p} \sum_{j=1}^p \mathbf{S}_j^{(r)}$ and broadcasts $\mathbf{S}^{(r)}$ to all devices.
 14: **if HEAPRIX variant:**
 15: Device j computes $\Phi_j^{(r)} \triangleq \text{HEAPRIX}[\mathbf{S}_j^{(r)}]$
 16: Second round of communication to obtain $\delta_j^{(r)} := \mathbf{S}_j \left(\text{HEAVYMIX}[\mathbf{S}^{(r)}] \right)$
 17: Broadcasts $\tilde{\mathbf{S}}^{(r)} \triangleq \frac{1}{p} \sum_{j=1}^p \delta_j^{(r)}$ to devices
 18: **end parallel for**
 19: **end**
 20: **Output:** $\mathbf{x}^{(R-1)}$

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

202 Assumption 1 is common in stochastic optimization. We present our results for PL, convex and
 203 general non-convex objectives. The reference [19] show that PL condition implies strong convexity
 204 property with same module (PL objectives can also be non-convex, hence strong convexity does not
 205 imply PL condition necessarily).

206 4.1 Convergence of FEDSKETCH

207 We now focus on the homogeneous case where data is i.i.d. among local devices, and therefore , the
 208 stochastic local gradient of each worker is an unbiased estimator of the global gradient. We have:

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

212 **Theorem 1.** *Suppose Assumptions 1-2 hold. Given $0 < m \leq d$ and considering Algorithm 3 with*
 213 *sketch size $B = O\left(m \log\left(\frac{dR}{\delta}\right)\right)$ and $\gamma \geq k$, with probability $1 - \delta$ we have:*

214 *In the **non-convex** case, $\{\mathbf{x}^{(r)}\}_{r=0}^{R-1}$ satisfies $\frac{1}{R} \sum_{r=0}^{R-1} \|\nabla f(\mathbf{x}^{(r)})\|_2^2 \leq \epsilon$ if:*

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

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

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

218 • **FS-PRIVIX**, for $\eta = \frac{1}{2L(cd/mk+1)\tau\gamma}$: $R = O((d/mk+1)\kappa \log(1/\epsilon))$ and $\tau =$
 219 $O\left((d/m+1)/(d/m+k)\epsilon\right)$.

220 • **FS-HEAPRIX**, for $\eta = \frac{1}{2L((cd-m)/mk+1)\tau\gamma}$: $R = O(((d-m)/mk+1)\kappa \log(1/\epsilon))$ and $\tau =$
 221 $O\left(d/m/(((d/m-1)+k)\epsilon)\right)$.

222 In the **Convex** case, $\{\mathbf{x}^{(r)}\}_{r=0}^{\infty}$ satisfies $\mathbb{E}[f(\mathbf{x}^{(R-1)}) - f(\mathbf{x}^{(*)})] \leq \epsilon$ if we set:

223 • **FS-PRIVIX**, for $\eta = \frac{1}{2L(cd/mk+1)\tau\gamma}$: $R = O(L(1+d/mk)/\epsilon \log(1/\epsilon))$ and $\tau =$
 224 $O\left((d/m+1)^2/(k(d/mk+1)^2\epsilon^2)\right)$.

225 • **FS-HEAPRIX**, for $\eta = \frac{1}{2L((cd-m)/mk+1)\tau\gamma}$: $R = O(L(1+(d-m)/mk)/\epsilon \log(1/\epsilon))$ and $\tau =$
 226 $O\left((d/m)^2/(k([d-m]/mk+1)^2\epsilon^2)\right)$.

227 The bounds in Theorem 1 suggest that in homogeneous setting if we set $d = m$ (no compression),
 228 the number of communication rounds to achieve the ϵ error matches with the number of iterations
 229 required to achieve the same error under a centralized setting. Additionally, computational complexity
 230 scales down with number of sampled devices. To stress on the further impact of using sketching, we
 231 also compare our results with prior works in terms of total number of communicated bits per device
 232 as follows:

233 **Comparison with Ivkin et al. [17]** From privacy aspect, we note Ivkin et al. [17] requires for
 234 server to have access to exact values of top_m gradients, hence do not preserve privacy, whereas our
 235 schemes do not need those exact values. From communication cost point of view, for strongly convex
 236 objective and compared to Ivkin et al. [17], we improve the total communication per worker from
 237 $RB = O\left(\frac{d}{\epsilon} \log\left(\frac{d}{\delta\sqrt{\epsilon}} \max\left(\frac{d}{m}, \frac{1}{\sqrt{\epsilon}}\right)\right)\right)$ to

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

238 We note that while reducing communication cost, our scheme requires $\tau = O(d/m(k(\frac{d}{mk}+1)\epsilon)) > 1$,
 239 which scales down with the number of sampled devices, k . Moreover, unlike Ivkin et al. [17], we do
 240 not use bounded gradient assumption. Therefore, we obtain stronger result with weaker assumptions.
 241 Regarding general non-convex objectives, our result improves the total communication cost per
 242 worker in Ivkin et al. [17] from $RB = O\left(\max(\frac{1}{\epsilon^2}, \frac{d^2}{k^2\epsilon}) \log(\frac{d}{\delta} \max(\frac{1}{\epsilon^2}, \frac{d^2}{k^2\epsilon}))\right)$ for *only one device*
 243 to $RB = O(\frac{m}{\epsilon} \log(\frac{d}{\epsilon\delta}))$. We also highlight that we can obtain similar rates for Algorithm 3 in
 244 heterogeneous environment if we make the additional assumption of uniformly bounded gradient.

245 **Note:** Such improved communication cost over prior related works is due to joint exploitation of
 246 *sketching*, to reduce the dimension of communicated messages, and the use of *local updates*, to
 247 reduce the total number of communication rounds leading to a specific convergence error.

248 4.2 Convergence of FedSKETCHGATE

249 We start with bounded local variance assumption:

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

254 **Theorem 2.** *Suppose Assumptions 1 and 3 hold. Given $0 < m \leq d$, and considering*
 255 *FedSKETCHGATE in Algorithm 4 with sketch size $B = O(m \log(\frac{dR}{\delta}))$ and $\gamma \geq p$ with proba-*
 256 *bility $1 - \delta$ we have*

257 In the **non-convex** case, $\eta = \frac{1}{L\gamma} \sqrt{\frac{mp}{R\tau(cd)}}$, $\{\mathbf{x}^{(r)}\}_{r=0}^{\infty}$ satisfies $\frac{1}{R} \sum_{r=0}^{R-1} \|\nabla f(\mathbf{x}^{(r)})\|_2^2 \leq \epsilon$ if:

258 • **FS-PRIVIX:**

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

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

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

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

262 • **FS-HEAPRIX**, for $\eta = m/(2cLd\tau\gamma)$: $R = O((\frac{d}{m})\kappa \log(1/\epsilon))$ and $\tau = O(1/(p\epsilon))$.

263 In the **convex** case, $\{\mathbf{x}^{(r)}\}_{r \geq 0}$ satisfies $\mathbb{E}[f(\mathbf{x}^{(R-1)}) - f(\mathbf{x}^{(*)})] \leq \epsilon$ if:

264 • **FS-PRIVIX**, for $\eta = 1/(2L(cd/m + 1)\tau\gamma)$: $R = O(L(d/m + 1)\epsilon \log(1/\epsilon))$ and $\tau =$

265 $O(1/(p\epsilon^2))$.

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

267 Theorem 2 implies that the number of communication rounds and local updates are similar to the

268 corresponding quantities in homogeneous setting except for the non-convex case where the number

269 of communication rounds also depends on the compression rate.

270 These results are summarized in Table 2-3 of the Appendix.

271 4.3 Comparison with Prior Methods

272 Before comparing with prior works, we highlight that privacy is another purpose of using unbiased

273 sketching in addition to communication efficiency. Therefore, our main competing schemes are

274 distributed algorithms based on sketching. Nonetheless, for the sake of showing the effectiveness of

275 our algorithms, we also compare with prior non-sketching based distributed algorithms ([20; 3; 36;

276 13]) in Section B of Appendix.

277 **Comparison with Li et al. [26]**. Note that our convergence analysis does not rely on the bounded

278 gradient assumption. We also improve both the number of communication rounds R and the size of

279 transmitted bits B per communication round. Additionally, we highlight that, while [26] provides a

280 convergence analysis for convex objectives, our analysis holds for PL (thus strongly convex case),

281 general convex and general non-convex objectives.

282 **Comparison with Rothchild et al. [37]**. Due to gradient tracking, our algorithm tackles data

283 heterogeneity issue, while algorithms in Rothchild et al. [37] does not particularly. As a consequence,

284 in FedSKETCHGATE each device has to store an additional state vector compared to Rothchild

285 et al. [37]. Yet, as our method is built upon an unbiased compressor, server does not need to

286 store any additional error correction vector. The convergence results for both of two variants of

287 FetchSGD in Rothchild et al. [37] rely on the uniform bounded gradient assumption which may

288 not be applicable with L -smoothness assumption when data distribution is highly heterogeneous,

289 as in FL, see [21], while our bounds do not assume such boundedness. Besides, Theorem 1 [37]

290 assumes that *Contraction Holds* for the sequence of gradients which may not hold in practice, yet

291 based on this strong assumption, their total communication cost (RB) in order to achieve ϵ error is

292 $RB = O\left(m \max(\frac{1}{\epsilon^2}, \frac{d^2 - dm}{m^2\epsilon}) \log\left(\frac{d}{\delta} \max(\frac{1}{\epsilon^2}, \frac{d^2 - dm}{m^2\epsilon})\right)\right)$. For the sake of comparison we let the

293 compression ratio in Rothchild et al. [37] to be $\frac{m}{d}$. In contrast, without any extra assumptions, our

294 results in Theorem 2 for PRIVIX and HEAPRIX are respectively $RB = O(\frac{(d+m)}{\epsilon} \log(\frac{(\frac{d^2}{m}) + d}{\epsilon\delta}))$ and

295 $RB = O(\frac{d}{\epsilon} \log(\frac{d^2}{\epsilon m\delta}))$ which improves the total communication cost of Theorem 1 in Rothchild

296 et al. [37] under regimes such that $\frac{1}{\epsilon} \geq d$ or $d \gg m$. Theorem 2 in Rothchild et al. [37] is based the

297 *Sliding Window Heavy Hitters* assumption, which is similar to the gradient diversity assumption in Li

298 et al. [29]; Haddadpour and Mahdavi [12]. Under that assumption the total communication cost is

299 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

300 window of gradients. We improve this bound under weaker assumptions in a regime where $\frac{I^{2/3}}{\epsilon^2} \geq d$.

301 We also provide bounds for PL, convex and non-convex objectives contrary to Rothchild et al. [37].

302 Finally, we note that algorithms in Rothchild et al. [37] are using momentum at server. While we do

303 not use it explicitly, we can modify our algorithms to include momentum easily.

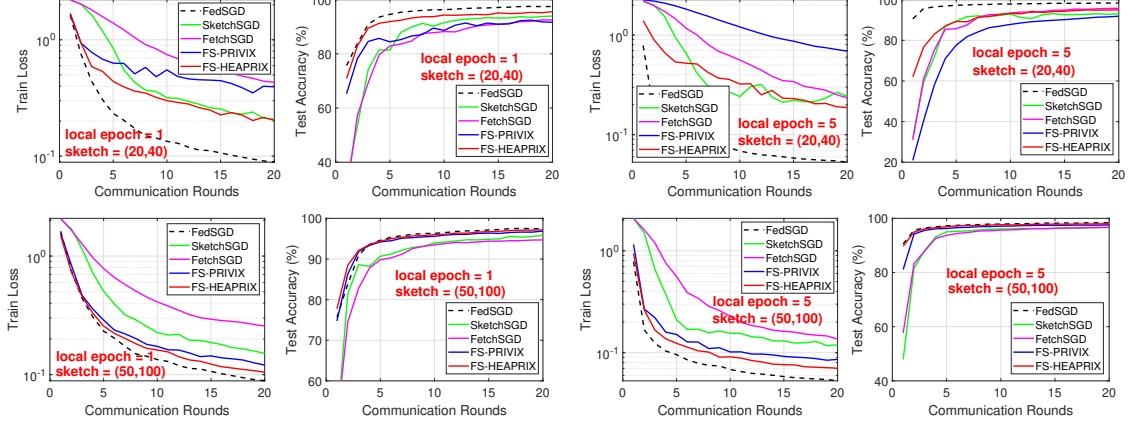


Figure 1: Homogeneous case: Comparison of compressed optimization methods on LeNet CNN.

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. [24], with 60 000 parameters. We compare Federated SGD (FedSGD) as the full-precision baseline, along with four sketching methods SketchSGD [17], FetchSGD [37], 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. [29]. 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 [37], 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.

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

- In general, increasing compression ratio would sacrifice learning performance. In all cases, FS-HEAPRIX performs the best in terms of both training objective 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., $75\times$).
- From the training loss, we see that the 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, our proposed 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,

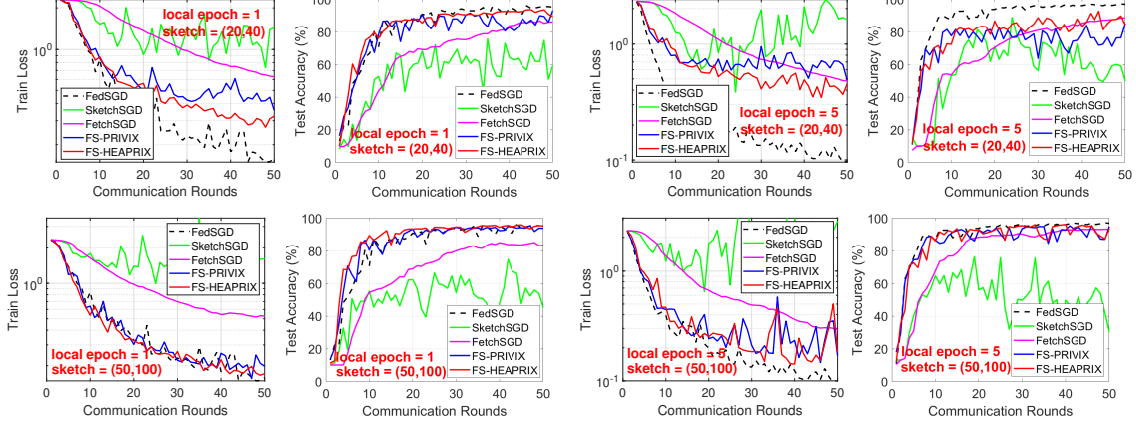


Figure 2: Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN.

341 which can achieve similar generalization accuracy as full-precision FedSGD, even with small sketch
 342 size (i.e., $75\times$ compression with 1 local epoch). Note that, slower convergence and worse generaliza-
 343 tion of FedSGD in non-i.i.d. data distribution case is also reported in e.g. McMahan et al. [33]; Chen
 344 et al. [8].

345 We also notice in Figure 2 the advantage of FS-HEAPRIX over FS-PRIVIX in terms of training
 346 loss and test accuracy. However, empirically we see that in the heterogeneous setting, more local
 347 updates tend to undermine the learning performance, especially with small sketch size. Nevertheless,
 348 when the sketch size is not too small, i.e., (50, 100), FS-HEAPRIX can still provide comparable test
 349 accuracy as FedSGD in both cases. Our empirical study demonstrates that our proposed FedSketch
 350 (and FedSketchGATE) frameworks are able to perform well in homogeneous (resp. heterogeneous)
 351 setting, with high compression rate. In particular, FedSketch methods are advantageous over recent
 352 SketchedSGD [17] and FetchSGD [37] in all cases. FS-HEAPRIX performs the best among all the
 353 tested compressed optimization algorithms, which in many cases achieves similar generalization
 354 accuracy as full-precision FedSGD with small sketch size.

355 6 Conclusion

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

References

- [1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720, Long Beach, 2017.
- [2] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5973–5983, Montréal, Canada, 2018.
- [3] Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-sgd: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14668–14679, Vancouver, Canada, 2019.
- [4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SIGNSGD: compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 559–568, Stockholmsmässan, Stockholm, Sweden, 2018.
- [5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1175–1191, Dallas, TX, 2017.
- [6] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 161–168, Vancouver, Canada, 2008.
- [7] Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004. doi: 10.1016/S0304-3975(03)00400-6. URL [https://doi.org/10.1016/S0304-3975\(03\)00400-6](https://doi.org/10.1016/S0304-3975(03)00400-6).
- [8] Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *ACM-IMS Foundations of Data Science Conference (FODS)*, Seattle, WA, 2020.
- [9] Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.
- [10] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.
- [11] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.
- [12] Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.
- [13] Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. *arXiv preprint arXiv:2007.01154*, 2020.
- [14] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677*, 2017.
- [15] Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*, 2020.
- [16] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

- 410 [17] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman
411 Arora. Communication-efficient distributed SGD with sketching. In *Advances in Neural
412 Information Processing Systems (NeurIPS)*, pages 13144–13154, Vancouver, Canada, 2019.
- 413 [18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Ar-
414 jun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings,
415 et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*,
416 2019.
- 417 [19] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-
418 gradient methods under the polyak-łojasiewicz condition. In *Proceedings of European Con-
419 ference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages
420 795–811, Riva del Garda, Italy, 2016.
- 421 [20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich,
422 and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated
423 learning. *arXiv preprint arXiv:1910.06378*, 2019.
- 424 [21] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on
425 identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence
426 and Statistics (AISTATS)*, pages 4519–4529, Online [Palermo, Sicily, Italy], 2020.
- 427 [22] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge
428 Discovery*, 7(4):373–397, 2003.
- 429 [23] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh,
430 and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv
431 preprint arXiv:1610.05492*, 2016.
- 432 [24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning
433 applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- 434 [25] Ping Li, Kenneth Ward Church, and Trevor Hastie. One sketch for all: Theory and application
435 of conditional random sampling. In *Advances in Neural Information Processing Systems (NIPS)*,
436 pages 953–960, Vancouver, Canada, 2008.
- 437 [26] Tian Li, Zaoxing Liu, Vyas Sekar, and Virginia Smith. Privacy for free: Communication-
438 efficient learning with differential privacy using sketches. *arXiv preprint arXiv:1911.00972*,
439 2019.
- 440 [27] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Chal-
441 lenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020.
- 442 [28] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Chal-
443 lenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.
- 444 [29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia
445 Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning
446 and Systems (MLSys)*, Austin, TX, 2020.
- 447 [30] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence
448 of fedavg on non-iid data. In *Proceedings of the 8th International Conference on Learning
449 Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.
- 450 [31] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Vari-
451 ance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*,
452 2019.
- 453 [32] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression:
454 Reducing the communication bandwidth for distributed training. In *Proceedings of the 6th
455 International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

- [33] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, Fort Lauderdale, FL, 2017.
- [34] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.
- [35] Constantin Philippenko and Aymeric Dieuleveut. Artemis: tight convergence guarantees for bidirectional compression in federated learning. *arXiv preprint arXiv:2006.14591*, 2020.
- [36] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2021–2031, Online [Palermo, Sicily, Italy], 2020.
- [37] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. FetchSGD: Communication-efficient federated learning with sketching. *arXiv preprint arXiv:2007.07682*, 2020.
- [38] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.
- [39] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.
- [40] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4447–4458, Montréal, Canada, 2018.
- [41] Sebastian Urban Stich. Local sgd converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.
- [42] Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7652–7662, Montréal, Canada, 2018.
- [43] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.
- [44] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems (NIPS)*, pages 1509–1519, Long Beach, CA, 2017.
- [45] Jiayang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized sgd and its applications to large-scale distributed optimization. *arXiv preprint arXiv:1806.08054*, 2018.
- [46] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7184–7193, Long Beach, CA, 2019.
- [47] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 5693–5700, Honolulu, HI, 2019.
- [48] Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3219–3227, Stockholm, Sweden, 2018.

505 A Notations and Definitions

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

Table 1: Table of Notations

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

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

512 A.1 Count sketch

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

Algorithm 5 Count Sketch (CS) [7]

```

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

```

514 A.2 PRIVIX and compression error of HEAPRIX

515 For the sake of completeness we review PRIVIX algorithm that is also mentioned in Li et al. [26] as
 516 follows:

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

```

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

```

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

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

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

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

B Summary of comparison of our results with prior works

For the purpose of further clarification, we summarize the comparison of our results with related works. We recall that p is the number of devices, 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. UG and PP stand for Unbounded Gradient and Privacy Property respectively.

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

531 **Comparison with Haddadpour et al. [13] and Reisizadeh et al. [36]** Convergence analysis of
 532 algorithms in [13] relies on unbiased compression, while in this paper our FL algorithm based on
 533 HEAPRIX enjoys from unbiased compression with equivalent biased compression variance. Moreover,
 534 we highlight that the convergence analysis of FedCOMGATE is based on the extra assumption of
 535 boundedness of the difference between the average of compressed vectors and compressed averages
 536 of vectors. However, we do not need this extra assumption as it is satisfied naturally due to linearity of
 537 sketching. Finally, as pointed out in Remark 2, our algorithms enjoy from a bidirectional compression
 538 property, unlike FedCOMGATE in general. Furthermore, since results in [13] improve the communica-
 539 tion complexity of FedPAQ algorithm, developed in [36], hence FedSKETCH and FedSKETCHGATE
 540 improves the communication complexity obtained in [36].

541 **Comparison with Basu et al. [3].** We note that the algorithm in [3] uses a composed compression
 542 and quantization while our algorithm is solely based on compression. So, in order to compare with
 543 algorithms in [3] we only consider Qsparse-local-SGD with compression and we let compression
 544 factor $\gamma = \frac{m}{d}$ (to compare with the same compression ratio induced with sketch size of mt). For
 545 strongly convex objective in Qsparse-local-SGD to achieve convergence error of ϵ they require $R =$
 546 $O\left(\kappa \frac{d}{m\sqrt{\epsilon}}\right)$ and $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$, which is improved to $R = O\left(\frac{\kappa d}{m} \log(1/\epsilon)\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$ for
 547 PL objectives. Similarly, for non-convex objective [3] requires $R = O\left(\frac{d}{m\epsilon^{1.5}}\right)$ and $\tau = O\left(\frac{m}{pd\sqrt{\epsilon}}\right)$,
 548 which is improved to $R = O\left(\frac{d}{m\epsilon}\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$. We note that we reduce communication
 549 rounds at the cost of increasing number of local updates (which scales down with number of
 550 devices, p). Additionally, we highlight that our FedSKETCHGATE exploits the gradient tracking
 551 idea to deal with data heterogeneity, while algorithms in [3] does not develop such mechanism
 552 and may suffer from poor convergence in heterogeneous setting. We also note that setting $\tau = 1$
 553 and using top_m compressor, the QSPARSE-local-SGD algorithm becomes similar to distributed
 554 SGD with sketching as they both use the error feedback framework to improve the compression
 555 variance. Finally, since the average of sparse vectors may not be sparse in general the number
 556 of transmitted bits from server to devices in QSPARSE-Local-SGD in [3] may not be sparse in
 557 general ($B = O(d)$), however our algorithms enjoy from bidirectional compression properly due to
 558 lower dimension and linearity properties of sketching ($B = O(m \log(\frac{Rd}{\delta}))$). Therefore, the total
 559 number of bits per device for strongly convex and non-convex objective is improved respectively from
 560 $RB = O\left(\kappa \frac{d^2}{m\sqrt{\epsilon}}\right)$ and $RB = O\left(\frac{d^2}{m\epsilon^{1.5}}\right)$ in [3] to $RB = O\left(\kappa d \log(\frac{\kappa d^2}{m\delta} \log(\frac{1}{\epsilon})) \log(1/\epsilon)\right) =$
 561 $O\left(\kappa d \max\left(\log(\frac{\kappa d^2}{m\delta}), \log^2(1/\epsilon)\right)\right)$ and $RB = O\left(\log(\frac{d^2}{m\epsilon\delta}) \frac{d}{\epsilon}\right)$.

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

565 **Comparison with Philippenko and Dieuleveut [35].** The reference [35] considers two-way com-
 566 pression from parameter server to devices and vice versa. They provide the convergence rate of
 567 $R = O\left(\frac{\omega^{\text{Up}} \omega^{\text{Down}}}{\epsilon^2}\right)$ for strongly-objective functions where ω^{Up} and ω^{Down} are uplink and downlink's
 568 compression noise (specializing to our case for the sake of comparison $\omega^{\text{Up}} = \omega^{\text{Down}} = \theta(d)$) for
 569 general heterogeneous data distribution. In contrast, while our algorithms are using bidirectional
 570 compression due to use of sketching for communication, our convergence rate for strongly-convex
 571 objective is $R = O(\kappa \mu^2 d \log(\frac{1}{\epsilon}))$ with probability $1 - \delta$.

572 C Theoretical Proofs

573 We will use the following fact (which is also used in Li et al. [30]; Haddadpour and Mahdavi [12]) in
 574 proving results.

575 **Fact 3** (Li et al. [30]; Haddadpour and Mahdavi [12]). *Let $\{x_i\}_{i=1}^p$ denote any fixed deterministic*
 576 *sequence. We sample a multiset \mathcal{P} (with size K) uniformly at random where x_j is sampled with*
 577 *probability q_j for $1 \leq j \leq p$ with replacement. Let $\mathcal{P} = \{i_1, \dots, i_K\} \subset [p]$ (some i_j s may have the*

578 same value). Then

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

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

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

584 We note that the sketching PRIVIX and HEAPRIX, satisfy Assumption 4 with $\omega = c \frac{d}{m}$ and $\omega =$
 585 $c \frac{d}{m} - 1$ respectively with probability $1 - \frac{\delta}{R}$ per communication round. Therefore, all the results in
 586 Theorem 1, by taking union over the all probabilities of each communication rounds, are concluded
 587 with probability $1 - \delta$ by plugging $\omega = c \frac{d}{m}$ and $\omega = c \frac{d}{m} - 1$ respectively into the corresponding
 588 convergence bounds.

589 C.1 Proof of Theorem 1

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

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

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

Proof.

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

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

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

595 Next we show that from Assumptions 3, we have

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

596 To do so, note that

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

597 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
598 in i.i.d. manner at each local machine, and ③ immediately follows from Assumptions 2.

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

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

600 Further note that we have

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

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

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

603 and the proof is complete. \square

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

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

606 *Proof.* We have:

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

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

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

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

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

611 *Proof.* Note that

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

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

612 where ① comes from $\mathbb{E}[\mathbf{x}^2] = \text{Var}[\mathbf{x}] + [\mathbb{E}[\mathbf{x}]]^2$ and ② holds because $\text{Var}\left(\sum_{j=1}^n \mathbf{x}_j\right) =$
613 $\sum_{j=1}^n \text{Var}(\mathbf{x}_j)$ for i.i.d. vectors \mathbf{x}_i (and i.i.d. assumption comes from i.i.d. sampling), and fi-
614 nally ③ follows from Assumption 2. \square

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

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

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

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

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

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

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

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

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

624 From the updating rule of Algorithm 3 we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

634 From the above inequality, is it easy to see that in order to achieve a linear speed up, we need to have

635 $\eta \gamma = O \left(\frac{\sqrt{k}}{\sqrt{R \tau}} \right).$ □

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

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

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

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

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

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

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

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

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

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

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

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

650 which matches the rate obtained in Wang and Joshi [43]. In this case the communication complexity
651 and the number of local updates become

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

652 which simply implies that in this special case the convergence rate of our algorithm reduces to the
653 rate obtained in Wang and Joshi [43], which indicates the tightness of our analysis.

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

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

657 **Theorem 5** (PL or strongly convex). For $\text{FedSKETCH}(\tau, \eta, \gamma)$, for all $0 \leq t \leq R\tau - 1$, under
658 Assumptions 1 to 2 and 3, if the learning rate satisfies

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

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

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

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

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

661 we obtain:

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

662 which leads to the following bound:

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

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

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

664

□

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

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

667 which indicates that to achieve an error of ϵ , we need to have $R = O \left(\left(\frac{\omega}{k} + 1 \right) \kappa \log \left(\frac{1}{\epsilon} \right) \right)$ and $\tau =$
668 $\frac{(\omega + 1)}{k(\frac{\omega}{k} + 1)\epsilon}$. Additionally, we note that if $\gamma \rightarrow \infty$, yet $R = O \left(\left(\frac{\omega}{k} + 1 \right) \kappa \log \left(\frac{1}{\epsilon} \right) \right)$ and $\tau = \frac{(\omega + 1)}{k(\frac{\omega}{k} + 1)\epsilon}$
669 will be necessary.

670 C.1.3 Main result for the general convex setting

671 **Theorem 6** (Convex). For a general convex function $f(\mathbf{w})$ with optimal solution $\mathbf{w}^{(*)}$, using
672 $\text{FedSKETCH}(\tau, \eta, \gamma)$ to optimize $\hat{f}(\mathbf{w}, \phi) = f(\mathbf{w}) + \frac{\phi}{2} \|\mathbf{w}\|^2$, for all $0 \leq t \leq R\tau - 1$, under
673 Assumptions 1 to 2, if the learning rate satisfies

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

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

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

675 We note that above theorem implies that to achieve a convergence error of ϵ we need to have
 676 $R = O\left(L\left(1 + \frac{\omega}{k}\right) \frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{(\omega+1)^2}{k(\frac{\omega}{k}+1)^2\epsilon}\right)$.

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

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

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

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

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

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

680 thus the proof is complete. \square

681 C.2 Proof of Theorem 2

682 The proof of Theorem 2 follows directly from the results in Haddadpour et al. [13]. We first mention
 683 the general Theorem 7 from [13] for general compression noise ω . Next, since the sketching PRIVIX
 684 and HEAPRIX, satisfy Assumption 4 with $\omega = c\frac{d}{m}$ and $\omega = c\frac{d}{m} - 1$ respectively with probability
 685 $1 - \frac{\delta}{R}$ per communication round, all the results in Theorem 2, conclude from Theorem 7 with
 686 probability $1 - \delta$ (by taking union over the all probabilities of each communication rounds with
 687 probability $1 - \delta/R$) and plugging $\omega = c\frac{d}{m}$ and $\omega = c\frac{d}{m} - 1$ respectively into the corresponding
 688 convergence bounds. For the heterogeneous setting, the results in Haddadpour et al. [13] requires the
 689 following extra assumption that naturally holds for the sketching:

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

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

695 Next, we restate the Theorem in Haddadpour et al. [13] here as follows:

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

- 699 • **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*
 700 *iterates satisfy $\frac{1}{R} \sum_{r=0}^{R-1} \|\nabla f(\mathbf{w}^{(r)})\|_2^2 \leq \epsilon$ if we set $R = O\left(\frac{\omega+1}{\epsilon}\right)$ and $\tau = O\left(\frac{1}{p\epsilon}\right)$.*
- 701 • **Strongly convex or PL:** *By choosing stepsizes as $\eta = \frac{1}{2L(\frac{\omega}{p}+1)\tau\gamma}$ and $\gamma \geq \sqrt{p\tau}$, we obtain*
 702 *that the iterates satisfy $\mathbb{E}[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})] \leq \epsilon$ if we set $R = O\left((\omega+1)\kappa \log\left(\frac{1}{\epsilon}\right)\right)$*
 703 *and $\tau = O\left(\frac{1}{p\epsilon}\right)$.*
- 704 • **Convex:** *By choosing stepsizes as $\eta = \frac{1}{2L(\omega+1)\tau\gamma}$ and $\gamma \geq \sqrt{p\tau}$, we obtain that the iterates*
 705 *satisfy $\mathbb{E}[f(\mathbf{w}^{(R)}) - f(\mathbf{w}^{(*)})] \leq \epsilon$ if we set $R = O\left(\frac{L(1+\omega)}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$ and $\tau = O\left(\frac{1}{p\epsilon^2}\right)$.*

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

710 D Numerical Experiments and Additional Results

711 D.1 Implementation of FetchSGD

712 Our implementation of FetchSGD basically follows the original paper (Algorithm 1 in [37]). The
 713 only difference is that, in the original algorithm, the local workers compress the gradient (in every
 714 local step) and transmit it to the central server. In our setting, we extend to the case with multiple local
 715 updates, where the difference in local weights are transmitted (same as the standard FL framework).
 716 Also, TopK compression is used to decode the sketches at the central server. We apply the same

717 implementation trick that when accumulating the errors, we only count the non-zero coordinates and
718 leave other coordinates zero for the accumulator. This greatly improves the empirical performance.

719 D.2 Additional Plots for the MNIST Experiments

720 D.2.1 Homogeneous setting

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

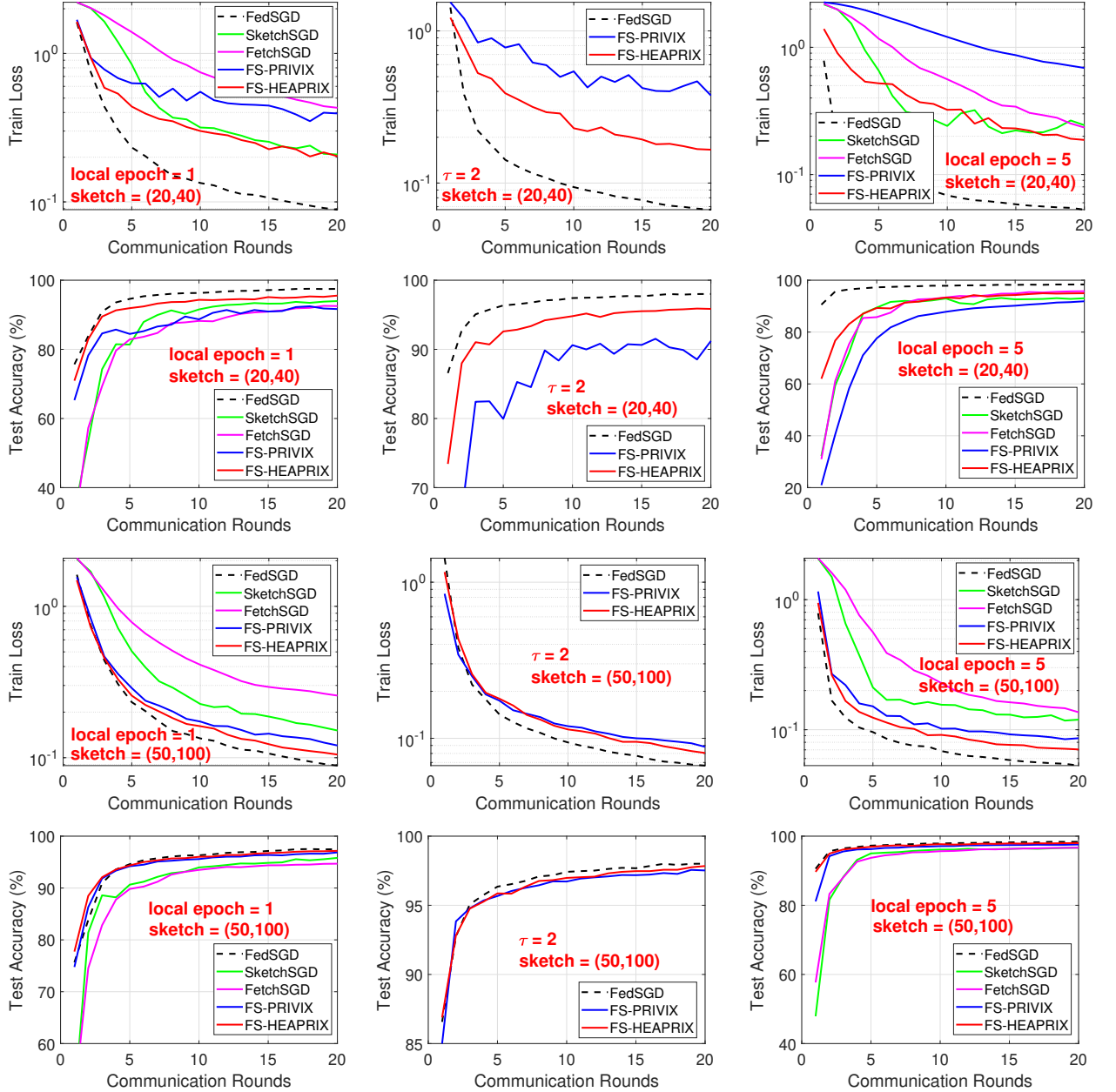


Figure 3: MNIST Homogeneous case: Comparison of compressed optimization methods on LeNet CNN architecture.

728 D.2.2 Heterogeneous setting

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

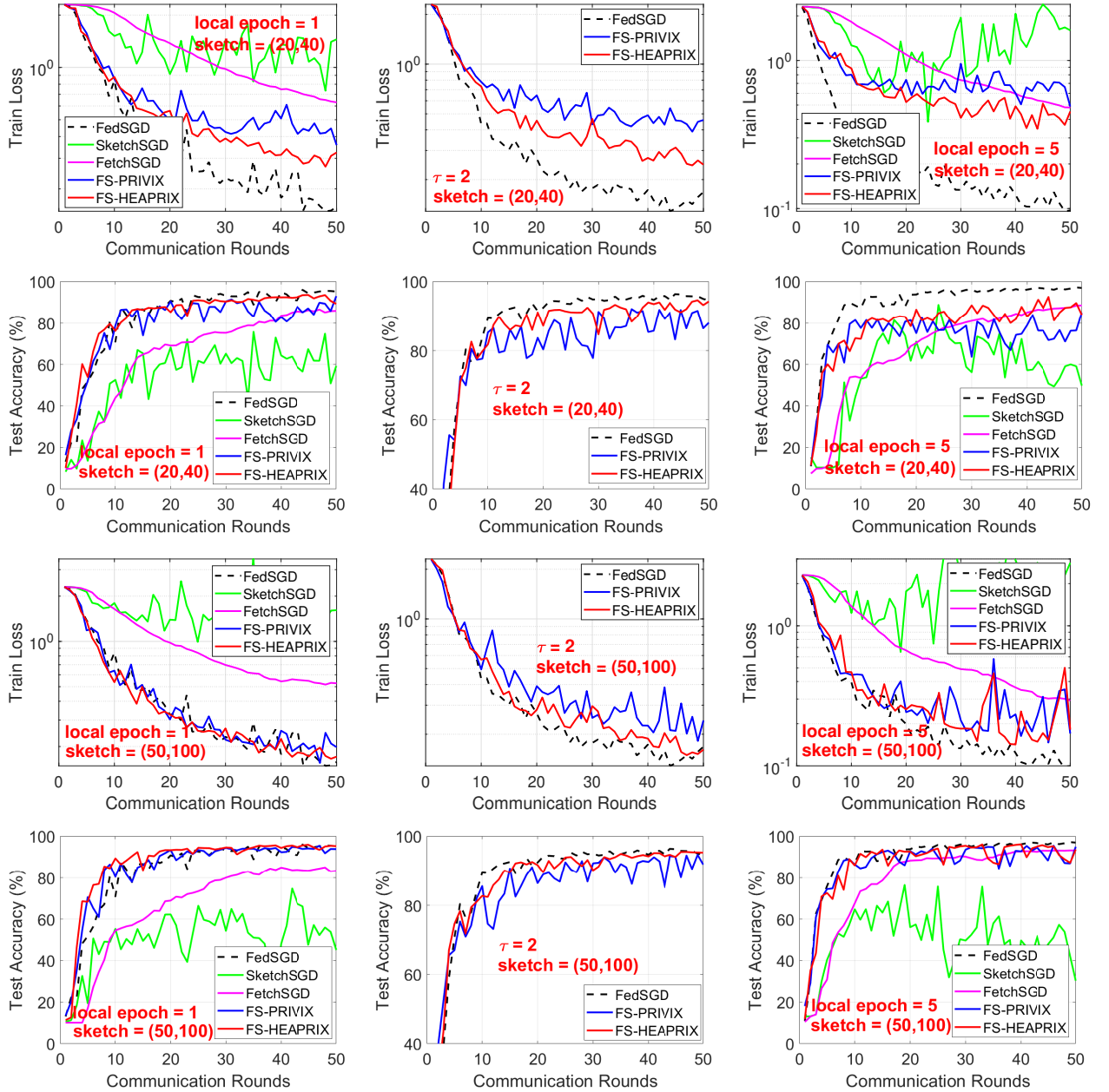


Figure 4: MNIST Heterogeneous case: Comparison of compressed optimization algorithms on LeNet CNN architecture.

733 D.3 Additional Experiments: CIFAR-10

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

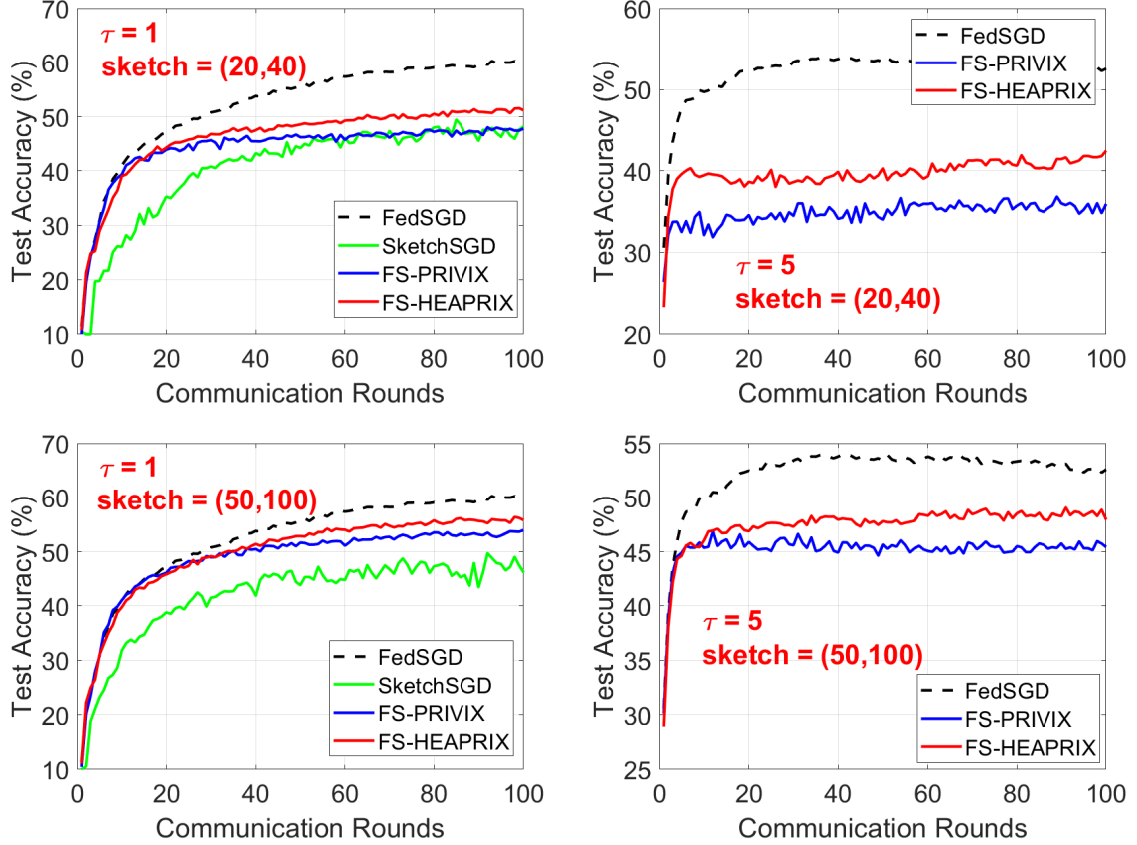


Figure 5: Homogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.

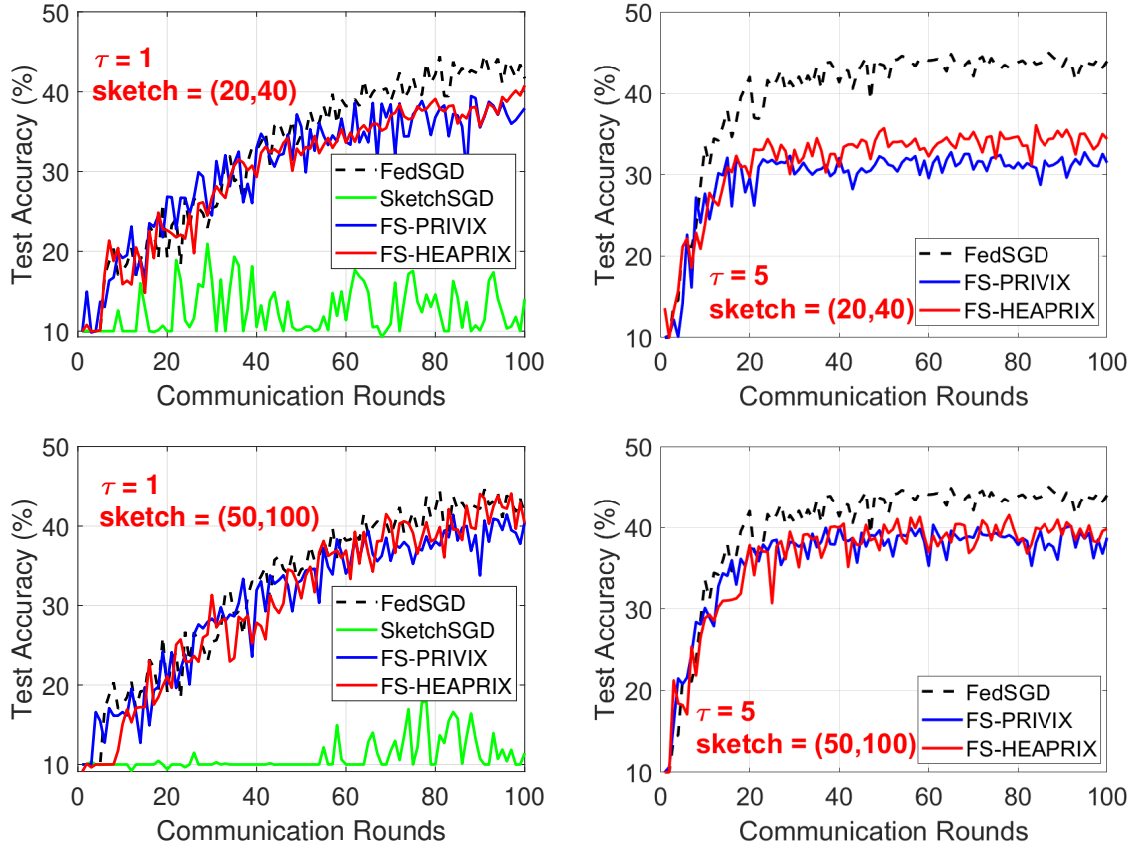


Figure 6: Heterogeneous case: CIFAR10: Comparison of compressed optimization methods on LeNet CNN.