

Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations

Debraj Basu*, Deepesh Data†, Can Karakus‡, Suhas Diggavi§

Abstract

Communication bottleneck has been identified as a significant issue in distributed optimization of large-scale learning models. Recently, several approaches to mitigate this problem have been proposed, including different forms of gradient compression or computing local models and mixing them iteratively. In this paper, we propose *Qsparse-local-SGD* algorithm, which combines aggressive sparsification with quantization and local computation along with error compensation, by keeping track of the difference between the true and compressed gradients. We propose both synchronous and asynchronous implementations of *Qsparse-local-SGD*. We analyze convergence for *Qsparse-local-SGD* in the *distributed* setting for smooth non-convex and convex objective functions. We demonstrate that *Qsparse-local-SGD* converges at the same rate as vanilla distributed SGD for many important classes of sparsifiers and quantizers. We use *Qsparse-local-SGD* to train ResNet-50 on ImageNet and show that it results in significant savings over the state-of-the-art, in the number of bits transmitted to reach target accuracy.

Keywords: Distributed optimization and learning; stochastic optimization; communication efficient training methods.

1 Introduction

Stochastic Gradient Descent (SGD) [HM51] and its many variants have become the workhorse for modern large-scale optimization as applied to machine learning [Bot10, BM11]. We consider a setup, in which SGD is applied to the *distributed* setting, where R different nodes compute *local* stochastic gradients on their *own* datasets \mathcal{D}_r . Co-ordination between them is done by aggregating these local computations to update the overall parameter \mathbf{x}_t as,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta_t}{R} \sum_{r=1}^R g_t^r,$$

where g_t^r , for $r = 1, 2, \dots, R$, is the local stochastic gradient at the r 'th machine for a local loss function $f^{(r)}(\mathbf{x})$ of the parameter vector \mathbf{x} , where $f^{(r)} : \mathbb{R}^d \rightarrow \mathbb{R}$ and η_t is the learning rate.

Training of high dimensional models is typically performed at a large scale over bandwidth limited networks. Therefore, despite the distributed processing gains, it is well understood by now that exchange of full-precision gradients between nodes causes communication to be the bottleneck for many large-scale models [AHJ⁺18, WXY⁺17, BWAA18, SYKM17]. For example,

*Adobe Inc.; dbasu@adobe.com; work done while at UCLA.

†UCLA; deepesh.data@gmail.com

‡Amazon Web Services; cakararak@amazon.com; work done while at UCLA.

§UCLA; suhasdiggavi@ucla.edu

consider training the BERT architecture for language models [DCLT18] which has about 340 million parameters, implying that each full precision exchange between workers is over 1.3GB. Such a communication bottleneck could be significant in emerging edge computation architectures suggested by federated learning [Kon17,MMR⁺17,ABC⁺16]. In such an architecture, data resides on and can even be generated by personal devices such as smart phones, and other edge (IoT) devices, in contrast to data-center architectures. Learning is envisaged with such an ultra-large scale, heterogeneous environment, with potentially unreliable or limited communication. These and other applications have led to many recently proposed methods, which are broadly based on three major approaches:

1. *Quantization* of gradients, where nodes locally quantize the gradient (perhaps with randomization) to a small number of bits [AGL⁺17,BWAA18,WHHZ18,WXY⁺17,SYKM17].
2. *Sparsification* of gradients, *e.g.*, where nodes locally select Top_k values of the gradient in absolute value and transmit these at full precision [Str15,AH17,SCJ18,AHJ⁺18,WHHZ18,LHM⁺18], while maintaining errors in local nodes for later compensation.
3. *Skipping communication rounds*, whereby nodes average their models after locally updating their models for several steps [YYZ19,Cop15,ZDW13,Sti19,CH16,WJ18].

In this work we propose a *Qsparse-local-SGD* algorithm, which combines aggressive sparsification with quantization and local computations, along with error compensation, by keeping track of the difference between the true and compressed gradients. We propose both synchronous and asynchronous implementations of *Qsparse-local-SGD* in a *distributed* setting, where the nodes perform computations on their local datasets. In our asynchronous model, the distributed nodes' iterates evolve at the same rate, but update the gradients at arbitrary times; see Section 4 for more details. We analyze convergence for *Qsparse-local-SGD* in the *distributed* case, for smooth non-convex and smooth strongly-convex objective functions. We demonstrate that *Qsparse-local-SGD* converges at the same rate as vanilla distributed SGD for many important classes of sparsifiers and quantizers. We implement *Qsparse-local-SGD* for ResNet-50 using the ImageNet dataset, and for a softmax multiclass classifier using the MNIST dataset, and we achieve target accuracies with about a factor of 15-20 savings over the state-of-the-art [AHJ⁺18,SCJ18,Sti19], in the total number of bits transmitted.

1.1 Related Work

The use of quantization for communication efficient gradient methods has decades rich history [GMT73] and its recent use in training deep neural networks [SFD⁺14,Str15] has re-ignited interest. Theoretically justified gradient compression using unbiased stochastic quantizers has been proposed and analyzed in [AGL⁺17,WXY⁺17,SYKM17]. Though methods in [WWLZ18,WSL⁺18] use induced sparsity in the quantized gradients, explicitly sparsifying the gradients more aggressively by retaining Top_k components, *e.g.*, $k < 1\%$, has been proposed [Str15,AH17,LHM⁺18,AHJ⁺18,SCJ18], combined with error compensation to ensure that all co-ordinates do get eventually updated as needed. [WHHZ18] analyzed error compensation for QSGD, without Top_k sparsification while focusing on quadratic functions. Another approach for mitigating the communication bottlenecks is by having infrequent communication, which has been popularly referred to in the literature as *iterative parameter mixing*, see [Cop15], and *model averaging*, see [Sti19,YYZ19,ZSMR16] and references therein. Our work is most closely related to and builds on the recent theoretical results in [AHJ⁺18,SCJ18,Sti19,YYZ19]. The analysis for the centralized Top_k (among other sparsifiers) was considered in [SCJ18], and [AHJ⁺18] analyzed a distributed version with the assumption of closeness of the aggregated Top_k gradients to the

centralized Top_k case, see Assumption 1 in [AHJ⁺18]. Local-SGD, where several local iterations are done before sending the *full* gradients, was studied in [Sti19, YYZ19], without any gradient compression beyond local iterations. Our work generalizes these works in several ways. We prove convergence for the *distributed* sparsification and error compensation algorithm, without the assumption of [AHJ⁺18], by using the perturbed iterate methods [MPP⁺17, SCJ18]. We analyze non-convex as well as convex objectives for the distributed case with local computations. A proof of sparsified SGD for convex objective functions and for the *centralized* case, without local computations¹ was given in [SCJ18]. Our techniques compose a (stochastic or deterministic 1-bit sign) quantizer with sparsification and local computations using error compensation. While our focus has only been on mitigating the communication bottlenecks in training high dimensional models over bandwidth limited networks, this technique works for any compression operator satisfying a regularity condition (see Definition 3) including our composed operators.

1.2 Contributions

We study a distributed set of R worker nodes, each of which perform computations on locally stored data, denoted by \mathcal{D}_r . Consider the empirical-risk minimization of the loss function

$$f(\mathbf{x}) = \frac{1}{R} \sum_{r=1}^R f^{(r)}(\mathbf{x})$$

where $f^{(r)}(\mathbf{x}) = \mathbb{E}_{i \sim \mathcal{D}_r} [f_i(\mathbf{x})]$, where $\mathbb{E}_{i \sim \mathcal{D}_r} [\cdot]$ denotes expectation over a random sample chosen from the local data set \mathcal{D}_r . Our setup can also handle different local functional forms, beyond dependence on the local data set \mathcal{D}_r , which is not explicitly written for notational simplicity. For $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote $\mathbf{x}^* := \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ and $f^* := f(\mathbf{x}^*)$. The distributed nodes perform computations and provide updates to the master node that is responsible for aggregation and model update. We develop *Qsparse-local-SGD*, a distributed SGD composing gradient quantization and explicit sparsification (e.g., Top_k components), along with local iterations. We develop the algorithms and analysis for both synchronous as well as asynchronous operations, in which workers can communicate with the master at arbitrary time intervals. To the best of our knowledge, these are the first algorithms which combine quantization, aggressive sparsification, and local computations for distributed optimization. With some minor modifications to *Qsparse-local-SGD*, it can also be used in a peer-to-peer setting, where the aggregation is done without any help from the master node, and each worker exchanges its updates with all other workers.

Our main theoretical results are the convergence analyses of *Qsparse-local-SGD* for both non-convex as well as convex objectives; see Theorem 1 and Theorem 3 for the synchronous case, as well as Theorem 4 and Theorem 6, for the asynchronous operation. Our analyses also demonstrate natural gains in convergence that distributed, mini-batch operation affords, and has convergence similar to equivalent vanilla SGD with local iterations (see Corollary 2 and Corollary 3), for both the non-convex case (with convergence rate $\sim \frac{1}{\sqrt{T}}$ for fixed learning rate) as well as the convex case (with convergence rate $\sim \frac{1}{T}$, for diminishing learning rate). We also demonstrate that quantizing and sparsifying the gradient, even after local iterations asymptotically yields an almost “free” efficiency gain (also observed numerically in Section 5 non-asymptotically). The numerical results on ImageNet dataset implemented for a ResNet-50

¹At the completion of our work, we recently found that in parallel to our work [KRSJ19] examined use of sign-SGD quantization, *without sparsification* for the centralized model. Another recent work in [KSJ19] studies the decentralized case with sparsification for strongly convex functions. In contrast, our work, developed independent of these works, uses quantization, sparsification and local computations for the distributed case, for both non-convex and strongly convex objectives.

architecture and for the convex case for multi-class logistic classification on MNIST [LBBH98] dataset demonstrates that one can get significant communication savings, while retaining equivalent state-of-the-art performance. The combination of quantization, sparsification, and local computations poses several challenges for theoretical analyses, including the analyses of impact of local iterations (block updates) of parameters on quantization and sparsification (see Lemma 4-5 in Section 3), as well as asynchronous updates and its combination with distributed compression (see Lemma 9-12 in Section 4).

1.3 Paper Organization

In Section 2, we demonstrate that composing certain classes of quantizers with sparsifiers satisfies a certain regularity condition that is needed for several convergence proofs for our algorithms. We describe the synchronous implementation of *Qsparse-local-SGD* in Section 3, and outline the main convergence results for it in Section 3.3, briefly giving the proof ideas in Section 3.4. We describe our asynchronous implementation of *Qsparse-local-SGD* and provide the theoretical convergence results in Section 4. The experimental results are given in Section 5. Many of the proof details are given in the appendices, given as part of the supplementary material.

2 Communication Efficient Operators

Traditionally, distributed stochastic gradient descent affords to send full precision (32 or 64 bit) unbiased gradient updates across workers to peers or to a central server that helps with aggregation. However, communication bottlenecks that arise in bandwidth limited networks limit the applicability of such an algorithm at a large scale when the parameter size is massive or the data is widely distributed on a very large number of worker nodes. In such settings, one could think of updates which not only result in convergence, but also require less bandwidth thus making the training process faster. In the following sections we discuss several useful operators from literature and enhance their use by proposing a novel class of composed operators.

We first consider two different techniques used in the literature for mitigating the communication bottleneck in distributed optimization, namely, quantization and sparsification. In quantization, we reduce precision of the gradient vector by mapping each of its components by a deterministic [BWAA18, KRSJ19] or randomized [AGL⁺17, WXY⁺17, SYKM17, ZDJW13] map to a finite number of quantization levels. In sparsification, we sparsify the gradients vector before using it to update the parameter vector, by taking its top k components, denoted by Top_k , or choosing k components uniformly at random, denoted by Rand_k , [SCJ18, KSJ19].

2.1 Quantization

SGD computes an unbiased estimate of the gradient, which can be used to update the model iteratively and is extremely useful in large scale applications. It is well known that the first order terms in the rate of convergence are affected by the variance of the gradients. While stochastic quantization of gradients could result in a variance blow up, it preserves the unbiasedness of the gradients at low precision; and, therefore, when training over bandwidth limited networks, the convergence would be much faster; see [AGL⁺17, WXY⁺17, SYKM17, ZDJW13].

Definition 1 (Randomized quantizer). *We say that $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a randomized quantizer with s quantization levels, if the following holds for every $\mathbf{x} \in \mathbb{R}^d$: (i) $\mathbb{E}_Q[Q_s(\mathbf{x})] = \mathbf{x}$; (ii) $\mathbb{E}_Q[\|Q_s(\mathbf{x})\|^2] \leq (1 + \beta_{d,s})\|\mathbf{x}\|^2$, where $\beta_{d,s} > 0$ could be a function of d and s . Here expectation is taken over the randomness of Q_s .*

Examples of randomized quantizers include

1. *QSGD* [AGL⁺17, WXY⁺17], which independently quantizes components of $\mathbf{x} \in \mathbb{R}^d$ into s levels, with $\beta_{d,s} = \min(\frac{d}{s^2}, \frac{\sqrt{d}}{s})$.
2. *Stochastic s -level Quantization* [SYKM17, ZDJW13], which independently quantizes every component of $\mathbf{x} \in \mathbb{R}^d$ into s levels between $\operatorname{argmax}_i x_i$ and $\operatorname{argmin}_i x_i$, with $\beta_{d,s} = \frac{d}{2s^2}$.
3. *Stochastic Rotated Quantization* [SYKM17], which is a stochastic quantization, preprocessed by a random rotation, with $\beta_{d,s} = \frac{2 \log_2(2d)}{s^2}$.

Instead of quantizing randomly into s levels, we can take a deterministic approach and round off each component of the vector to the nearest level. In particular, we can just take the sign, which has shown promise in [BWAA18, KRSJ19].

Definition 2 (Deterministic Sign quantizer). *A deterministic quantizer $\operatorname{Sign} : \mathbb{R}^d \rightarrow \{+1, -1\}^d$ is defined as follows: for every vector $\mathbf{x} \in \mathbb{R}^d$, the i 'th component of $\operatorname{Sign}(\mathbf{x})$, for $i \in [d]$, is defined as $\mathbb{1}\{x_i \geq 0\} - \mathbb{1}\{x_i < 0\}$.*

Such methods drew interest since RPROP [RB93], which only used the temporal behavior of the sign of the gradient. This is an example where the biased 1-bit quantizer as in Definition 2 is used. This further inspired optimizers, such as RMSPROP [TH12], ADAM [KB15], which incorporate appropriate adaptive scaling with momentum acceleration and have demonstrated empirical superiority in non-convex applications.

2.2 Sparsification

As mentioned earlier, we consider two important examples of sparsification operators: Top_k and Rand_k . For any $\mathbf{x} \in \mathbb{R}^d$, $\operatorname{Top}_k(\mathbf{x})$ is equal to a d -length vector, which has at most k non-zero components whose indices correspond to the indices of the largest k components (in absolute value) of \mathbf{x} . Similarly, $\operatorname{Rand}_k(\mathbf{x})$ is a d -length (random) vector, which is obtained by selecting k components of \mathbf{x} uniformly at random. Both of these satisfy a so-called ‘‘compression’’ property as defined below, with $\gamma = k/d$ [SCJ18].

Definition 3 (Compression operator [SCJ18]). *A (randomized) function $\operatorname{Comp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called a compression operator, if there exists a constant $\gamma \in (0, 1]$ (that may depend on k and d), such that for every $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}_C[\|\mathbf{x} - \operatorname{Comp}_k(\mathbf{x})\|_2^2] \leq (1 - \gamma)\|\mathbf{x}\|_2^2, \quad (1)$$

where expectation is taken over the randomness of the compression operator Comp_k .

Note that stochastic quantizers, as defined in Definition 1, also satisfy this regularity condition in Definition 3 for $\beta_{d,s} \leq 1$. Now we give a simple but important corollary, which allows us to apply different compression operators to different coordinates of a vector. As an application, in the case of training neural networks, we can apply different operators to different layers.

Corollary 1 (Piecewise compression). *Let $C_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$ for $i \in [L]$ denote possibly different compression operators with compression coefficients γ_i . Let $\mathbf{x} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_L]$, where $\mathbf{x}_i \in \mathbb{R}^{d_i}$ for all $i \in [L]$. Then $C(\mathbf{x}) := [C_1(\mathbf{x}_1) C_2(\mathbf{x}_2) \dots C_L(\mathbf{x}_L)]$ is a compression operator with the compression coefficient being equal to $\gamma_{\min} = \min_{i \in [L]} \gamma_i$.*

Proof. Fix an arbitrary $\mathbf{x} \in \mathbb{R}^d$. The result follows from the following set of inequalities: $\mathbb{E}_C\|\mathbf{x} - C(\mathbf{x})\|_2^2 = \sum_{i=1}^L \mathbb{E}_{C_i}\|\mathbf{x}_i - C_i(\mathbf{x}_i)\|_2^2 \stackrel{(a)}{\leq} \sum_{i=1}^L (1 - \gamma_i)\|\mathbf{x}_i\|_2^2 \leq (1 - \gamma_{\min})\|\mathbf{x}\|_2^2$, where inequality (a) follows because each C_i is a compression operator with the compression coefficient γ_i . \square

Corollary 1 allows us to apply different compression operators to different coordinates of the updates which can be based upon their dimensionality and sparsity patterns.

2.3 Composition of Quantization and Sparsification

Now we show that we can compose deterministic/randomized quantizers with sparsifiers and the resulting operator is a compression operator. First we compose a general stochastic quantizer with an explicit sparsifier, such as $\text{Top}_k(\mathbf{x})$ and $\text{Rand}_k(\mathbf{x})$, and show that the resulting operator is a ‘‘compression’’ operator. A proof is provided in [Appendix A.1](#).

Lemma 1 (Compression of a composed operator). *Let $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a quantizer with parameter s that satisfies [Definition 1](#). Let $Q_s \text{Comp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined as $Q_s \text{Comp}_k(\mathbf{x}) := Q_s(\text{Comp}_k(\mathbf{x}))$ for every $\mathbf{x} \in \mathbb{R}^d$. If k, s are such that $\beta_{k,s} < 1$, then $Q_s \text{Comp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a compression operator with the compression coefficient being equal to $\gamma = (1 - \beta_{k,s}) \frac{k}{d}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \leq \left[1 - (1 - \beta_{k,s}) \frac{k}{d}\right] \|\mathbf{x}\|_2^2,$$

where expectation is taken over the randomness of the compression operator Comp_k as well as of the quantizer Q_s .

For the different quantizers mentioned earlier, the conditions when their composition with Comp_k gives $\beta_{k,s} < 1$ are:

1. *QSGD*: for $k < s^2$, we get. $\gamma = (1 - \frac{k}{s^2}) \frac{k}{d}$
2. *Stochastic k -level Quantization*: for $k < 2s^2$, we get $\gamma = (1 - \frac{k}{2s^2}) \frac{k}{d}$.
3. *Stochastic Rotated Quantization*: for $k < 2s^{2/2-1}$, we get $\gamma = (1 - \frac{2 \log_2(2k)}{s^2}) \frac{k}{d}$.

Remark 1. *Observe that for a given stochastic quantizer that satisfies [Definition 1](#), we have a prescribed operating regime of $\beta_{k,s} < 1$. This results in an upper bound on the coarseness of the quantizer, which happens because the quantization leads to a blow-up of the second moment; see condition (ii) of [Definition 1](#). However, by employing [Corollary 1](#), we show that this can be alleviated to some extent via an example.*

Consider an operator as described in [Lemma 1](#), where the quantizer, $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ in use is QSGD [[AGL⁺17](#), [WXY⁺17](#)], and the sparsifier, Comp_k is Top_k [[AHJ⁺18](#), [SCJ18](#)]. Apply it to a vector $\mathbf{x} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_L] \in \mathbb{R}^d$ in a piecewise manner, i.e., $Q_{s_i} \text{Comp}_{k_i} : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$ to smaller vectors $\mathbf{x}_i \in \mathbb{R}^{d_i}$ as prescribed in [Corollary 1](#). Define $\beta_{k_i, s_i} = \frac{k_i}{s_i^2}$ as the coefficient of the variance bound as in [Definition 1](#) for the quantizer Q_{s_i} , used for \mathbf{x}_i and $k := \sum_{i=1}^L k_i$. Observe that the regularity condition in [Definition 3](#) can be satisfied by having $k_i < s_i^2$. Therefore, the piecewise compression operator allows a coarser quantizer than when the operator is applied to the entire vector together where we require $\beta_{k,s} = \frac{k}{s^2} < 1$, thus providing a small gain in communication efficiency. For example, consider the composed operator being applied on a per layer basis to a deep neural network. We can now afford to have a much coarser quantizer than when the operator is applied to all the parameters at once.

As discussed above, stochastic quantization results in a variance blow-up which limits our regime of operation, when we combine that with sparsification. However, it turns out that, we can expand our regime of operation unrestrictedly by scaling the vector $Q_s \text{Comp}_k(\mathbf{x})$ appropriately. We summarize the result in the following lemma, which is proved in [Appendix A.2](#).

Lemma 2 (Composing sparsification with stochastic quantization). *Let $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a stochastic quantizer with parameter s that satisfies [Definition 1](#).*

Let $Q_s \text{Comp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined as $Q_s \text{Comp}_k(\mathbf{x}) := Q_s(\text{Comp}_k(\mathbf{x}))$ for every $\mathbf{x} \in \mathbb{R}^d$. Then $\frac{Q_s \text{Comp}_k(\mathbf{x})}{1+\beta_{k,s}}$ is a compression operator with the compression coefficient being equal to $\gamma = \frac{k}{d(1+\beta_{k,s})}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, we have

$$\mathbb{E}_{C,Q} \left[\left\| \mathbf{x} - \frac{Q_s \text{Comp}_k(\mathbf{x})}{1+\beta_{k,s}} \right\|_2^2 \right] \leq \left[1 - \frac{k}{d(1+\beta_{k,s})} \right] \|\mathbf{x}\|_2^2.$$

Remark 2. Note that, unlike $Q_s \text{Comp}_k(\mathbf{x})$, the scaled version $\frac{Q_s \text{Comp}_k(\mathbf{x})}{1+\beta_{k,s}}$ is always a compression operator for all values of $\beta_{k,s} > 0$. Furthermore, observe that, if $\beta_{k,s} < 1$, then we have $(1 - \beta_{k,s}) \frac{k}{d} < \frac{k}{d(1+\beta_{k,s})}$, which implies that even in the operating regime of $\beta_{k,s} < 1$, which is required in [Lemma 1](#), the scaled composed operator $\frac{Q_s \text{Comp}_k(\mathbf{x})}{1+\beta_{k,s}}$ of [Lemma 2](#) gives better compression than what we get from the unscaled composed operator $Q_s \text{Comp}_k(\mathbf{x})$ of [Lemma 1](#). So, appropriately scaled composed operator is always a better choice for compression.

We can also compose a *deterministic* 1-bit quantizer Sign with Comp_k . For that we need some notations first. For $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$ and given vector $\mathbf{x} \in \mathbb{R}^d$, let $\mathcal{S}_{\text{Comp}_k(\mathbf{x})} \in \binom{[d]}{k}$ denote the set of k indices chosen for defining $\text{Comp}_k(\mathbf{x})$. For example, if $\text{Comp}_k = \text{Top}_k$, then $\mathcal{S}_{\text{Comp}_k(\mathbf{x})}$ denote the set of k indices corresponding to the largest k components of \mathbf{x} ; if $\text{Comp}_k = \text{Rand}_k$, then $\mathcal{S}_{\text{Comp}_k(\mathbf{x})}$ denote a set of random set of k indices in $[d]$. The composition of Sign with $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$ is denoted by $\text{SignComp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$, and for $i \in [d]$, the i 'th component of $\text{SignComp}_k(\mathbf{x})$ is defined as

$$(\text{SignComp}_k(\mathbf{x}))_i := \begin{cases} \mathbb{1}\{x_i \geq 0\} - \mathbb{1}\{x_i < 0\} & \text{if } i \in \mathcal{S}_{\text{Comp}_k(\mathbf{x})}, \\ 0 & \text{otherwise.} \end{cases}$$

In the following lemma we show that SignComp_k is a compression operator; a proof of which is provided in [Appendix A.3](#).

Lemma 3 (Composing sparsification with deterministic quantization). *For $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$, the operator*

$$\frac{\|\text{Comp}_k(\mathbf{x})\|_m \text{SignComp}_k(\mathbf{x})}{k}$$

for any $m \in \mathbb{Z}_+$ is a compression operator with the compression coefficient γ_m being equal to

$$\gamma_m = \begin{cases} \max \left\{ \frac{1}{d}, \frac{k}{d} \left(\frac{\|\text{Comp}_k(\mathbf{x})\|_1}{\sqrt{d} \|\text{Comp}_k(\mathbf{x})\|_2} \right)^2 \right\} & \text{if } m = 1, \\ \frac{k^{\frac{2}{m}-1}}{d} & \text{if } m \geq 2. \end{cases}$$

Remark 3. Observe that for $m = 1$, depending on the value of k , either of the terms inside the max can be bigger than the other term. For example, if $k = 1$, then $\|\text{Comp}_k(\mathbf{x})\|_1 = \|\text{Comp}_k(\mathbf{x})\|_2$, which implies that the second term inside the max is equal to $1/d^2$, which is much smaller than the first term. On the other hand, if $k = d$ and the vector \mathbf{x} is dense, then the second term may be much bigger than the first term.

3 Distributed Synchronous Operation

Let $\mathcal{I}_T^{(r)} \subseteq [T] := \{1, \dots, T\}$ with $T \in \mathcal{I}_T^{(r)}$ denote a set of indices for which worker $r \in [R]$ synchronizes with the master. In a synchronous setting, $\mathcal{I}_T^{(r)}$ is same for all the workers. Let

$\mathcal{I}_T := \mathcal{I}_T^{(r)}$ for any $r \in [R]$. Every worker $r \in [R]$ maintains a local parameter vector $\widehat{\mathbf{x}}_t^{(r)}$ which is updated in each iteration t . If $t \in \mathcal{I}_T$, every worker $r \in [R]$ sends the compressed and error-compensated update $g_t^{(r)}$ computed on the net progress made since the last synchronization to the master node, and updates its local memory $m_t^{(r)}$. Upon receiving $g_t^{(r)}, r = 1, 2, \dots, R$, master aggregates them, updates the global parameter vector, and sends the new model \mathbf{x}_{t+1} to all the workers; upon receiving which, they set their local parameter vector $\widehat{\mathbf{x}}_{t+1}^{(r)}$ to be equal to the global parameter vector \mathbf{x}_{t+1} . Our algorithm is summarized in [Algorithm 1](#).

Algorithm 1 Qsparse-local-SGD

```

1: Initialize  $\mathbf{x}_0 = \widehat{\mathbf{x}}_0^{(r)} = m_0^{(r)} = \mathbf{0}, \forall r \in [R]$ . Suppose  $\eta_t$  follows a certain learning rate schedule.
2: for  $t = 0$  to  $T - 1$  do
3:   On Workers:
4:   for  $r = 1$  to  $R$  do
5:      $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} \leftarrow \widehat{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}(\widehat{\mathbf{x}}_t^{(r)})$ ;  $i_t^{(r)}$  is a mini-batch of size  $b$  uniformly in  $\mathcal{D}_r$ 
6:     if  $t + 1 \notin \mathcal{I}_T$  then
7:        $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t, m_{t+1}^{(r)} \leftarrow m_t^{(r)}$  and  $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$ 
8:     else
9:        $g_t^{(r)} \leftarrow QComp_k(m_t^{(r)} + \mathbf{x}_t - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)})$ , send  $g_t^{(r)}$  to the master
10:       $m_{t+1}^{(r)} \leftarrow m_t^{(r)} + \mathbf{x}_t - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} - g_t^{(r)}$ 
11:      Receive  $\mathbf{x}_{t+1}$  from the master and set  $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \mathbf{x}_{t+1}$ 
12:    end if
13:  end for
14:  At Master:
15:  if  $t + 1 \notin \mathcal{I}_T$  then
16:     $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$ 
17:  else
18:    Receive  $g_t^{(r)}$  from  $R$  workers and compute  $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{R} \sum_{r=1}^R g_t^{(r)}$ 
19:    Broadcast  $\mathbf{x}_{t+1}$  to all workers
20:  end if
21: end for
22: Comment:  $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$  is used to denote an intermediate variable between iterations  $t$  and  $t + 1$ 

```

3.1 Assumptions

All results in this paper use the following two standard assumptions.

1. **Smoothness:** The local function $f^{(r)} : \mathbb{R}^d \rightarrow \mathbb{R}$ at each worker $r \in [R]$ is L -smooth, i.e., for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $f^{(r)}(\mathbf{y}) \leq f^{(r)}(\mathbf{x}) + \langle \nabla f^{(r)}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$.
2. **Bounded second moment:** For every $\widehat{\mathbf{x}}_t^{(r)} \in \mathbb{R}^d, r \in [R], t \in [T]$ and for some constant $0 \leq G < \infty$, we have $\mathbb{E}_{i \sim \mathcal{D}_r} [\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)})\|_2^2] \leq G^2$. This is a standard assumption in [SSS07, NJLS09, RRWN11, HK14, RSS12, SCJ18, Sti19, YYZ19, KSJ19, AHJ⁺18]. Relaxation of the uniform boundedness of the gradient allowing arbitrarily different gradients of local functions in heterogenous settings as done for SGD in [NNvD⁺18, WJ18] is left for future work. This also imposes a **bound on the variance**: $\mathbb{E}_{i \sim \mathcal{D}_r} [\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)}) - \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|_2^2] \leq \sigma_r^2$, where $\sigma_r^2 \leq G^2$ for every $r \in [R]$.

In this section we present our main convergence results with synchronous updates, obtained by

running [Algorithm 1](#) for smooth functions, both non-convex and strongly convex. To state our results, we need the following definition from [\[Sti19\]](#).

Definition 4 (Gap [\[Sti19\]](#)). Let $\mathcal{I}_T = \{t_0, t_1, \dots, t_k\}$, where $t_i < t_{i+1}$ for $i = 0, 1, \dots, k-1$. The gap of \mathcal{I}_T is defined as $\text{gap}(\mathcal{I}_T) := \max_{i \in [k]} \{t_i - t_{i-1}\}$, which is equal to the maximum difference between any two consecutive synchronization indices.

3.2 Error Compensation

Sparsified gradient methods, where workers send the largest k coordinates of the updates based on their magnitudes have been investigated in the literature and serves as a communication efficient strategy for distributed training of learning models. However, the convergence rates are subpar to distributed vanilla SGD. Together with some form of error compensation, these methods have been empirically observed to converge as fast as vanilla SGD in [\[Str15, AH17, LHM⁺18, AHJ⁺18, SCJ18\]](#). In [\[AHJ⁺18, SCJ18\]](#), sparsified SGD with such feedback schemes has been carefully analyzed. Under analytic assumptions, [\[AHJ⁺18\]](#) proves the convergence of distributed Top_k SGD with error feedback. The net error in the system is accumulated by each worker locally on a per iteration basis and this is used as feedback for generating the future updates. [\[SCJ18\]](#) did the analysis for the centralized Top_k SGD for strongly convex objectives.

In [Algorithm 1](#), the error introduced in every iteration is accumulated into the memory of each worker, which is compensated for in the future rounds of communication. This feedback is the key to recovering the convergence rates matching vanilla SGD. The operators employed provide a controlled way of using both the current update as well as the compression errors from the previous rounds of communication. Under the assumption of the uniform boundedness of the gradients, we analyze the controlled evolution of memory through the optimization process; the results are summarized in [Lemma 4](#) and [Lemma 5](#) below.

3.2.1 Decaying Learning Rate

Here we show that if we run [Algorithm 1](#) with a decaying learning rate η_t , then the local memory at each worker contracts and goes to zero as $\mathcal{O}(\eta_t)^2$.

Lemma 4 (Memory contraction). Let $\text{gap}(\mathcal{I}_T) \leq H$ and $\eta_t = \frac{\xi}{a+t}$, where ξ is a constant and $a > \frac{4H}{\gamma}$, with γ being the compression coefficient of the compression operator. Then there exists a constant $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$, such that the following holds for every $t \in \mathbb{Z}^+$ and $r \in [R]$:

$$\mathbb{E}\|m_t^{(r)}\|_2^2 \leq 4 \frac{\eta_t^2}{\gamma^2} C H^2 G^2. \quad (2)$$

We prove [Lemma 4](#) in [Appendix B.1](#). Note that for fixed γ, H , the memory decays as $\mathcal{O}(\eta_t^2)$. This implies that the net error in the algorithm from the compression of updates in each round of communication is compensated for in the end.

3.2.2 Fixed Learning Rate

In the following lemma, which is proved in [Appendix B.2](#), we show that if we run [Algorithm 1](#) with a fixed learning rate $\eta_t = \eta, \forall t$, then the local memory at each worker is bounded. It can be verified that the proof of [Lemma 4](#) also holds for fixed learning rate, and we can trivially bound $\mathbb{E}\|m_t^{(r)}\|_2^2$ in this case by simply putting $\eta_t = \eta$ in (2). However, we can get a better bound (saving a factor of $\frac{C}{1-\gamma^2}$, which is bigger than 4) by directly working with a fixed learning rate.

Lemma 5 (Bounded memory). *Let $\text{gap}(\mathcal{I}_T) \leq H$. Then the following holds for every worker $r \in [R]$ and for every $t \in \mathbb{Z}^+$:*

$$\mathbb{E}\|m_t^{(r)}\|_2^2 \leq 4 \frac{\eta^2(1-\gamma^2)}{\gamma^2} H^2 G^2. \quad (3)$$

Note that, for fixed γ, H , the memory is upper bounded by a constant $\mathcal{O}(\eta^2)$. Observe that since the memory accumulates the past errors due to compression and local computation, in order to asymptotically reduce the memory to zero, the learning rate would have to be reduced once in a while throughout the training process.

3.3 Main Results

We leverage the perturbed iterate analysis as in [MPP⁺17, SCJ18] to provide convergence guarantees for *Qsparse-local-SGD*. Under the assumptions of Section 3.1, the following theorems hold when Algorithm 1 is run with any compression operator (including our composed operators).

Theorem 1 (Smooth (non-convex) case with fixed learning rate). *Let $f^{(r)}(\mathbf{x})$ be L -smooth for every $i \in [R]$. Let $QComp_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a compression operator whose compression coefficient is equal to $\gamma \in (0, 1]$. Let $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ be generated according to Algorithm 1 with $QComp_k$, for step sizes $\eta = \frac{\widehat{C}}{\sqrt{T}}$ (where \widehat{C} is a constant such that $\frac{\widehat{C}}{\sqrt{T}} \leq \frac{1}{2L}$) and $\text{gap}(\mathcal{I}_T) \leq H$. Then we have*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{\widehat{C}} + \widehat{C}L \left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \right) \right) \frac{4}{\sqrt{T}} + 8 \left(4 \frac{(1-\gamma^2)}{\gamma^2} + 1 \right) \frac{\widehat{C}^2 L^2 G^2 H^2}{T}.$$

Here \mathbf{z}_T is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability $1/RT$.

Corollary 2. *Let $\mathbb{E}[f(\mathbf{x}_0)] - f^* \leq J^2$, where $J < \infty$ is a constant,² $\sigma_{\max} = \max_{r \in [R]} \sigma_r$, and $\widehat{C}^2 = \frac{bR(\mathbb{E}[f(\mathbf{x}_0)] - f^*)}{\sigma_{\max}^2 L}$, we have*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \mathcal{O}\left(\frac{J\sigma_{\max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2 b R G^2 H^2}{\sigma_{\max}^2 \gamma^2 T}\right).$$

In order to ensure that the compression does not affect the dominating terms while converging at a rate of $\mathcal{O}\left(1/\sqrt{bRT}\right)$, we would require $H = \mathcal{O}\left(\gamma T^{1/4}/(bR)^{3/4}\right)$.

Theorem 1 is proved in Appendix B.6 and provides non-asymptotic guarantees, where we observe that compression does not affect the first order term. Here, we are required to decide the horizon T before running the algorithm. Therefore, in order to converge to a fixed point, the learning rate needs to follow a piecewise schedule (i.e., the learning rate would have to be reduced once in a while throughout the training process), which is also the case in our numerics in Section 5.1. The corresponding asymptotic result (with decaying learning rate) is given below.

Theorem 2 (Smooth (non-convex) case with decaying learning rate). *Let $f^{(r)}(\mathbf{x})$ be L -smooth for every $r \in [R]$. Let $QComp_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a compression operator whose compression coefficient is equal to $\gamma \in (0, 1]$. Let $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ be generated according to Algorithm 1 with $QComp_k$, for step sizes $\eta_t = \frac{\xi}{(a+t)}$ and $\text{gap}(\mathcal{I}_T) \leq H$, where $a > 1$ is such that we have $a > \max\left\{\frac{4H}{\gamma}, 2\xi L, H\right\}$ and $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$. Then the following holds.*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2}{(a-1)P_T} \left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \right) + \left(\frac{8C}{\gamma^2} + 8 \right) \frac{\xi^3 L^2 G^2 H^2}{2(a-1)^2 P_T}.$$

²Even classical SGD requires knowing an upper bound on $\|\mathbf{x}_0 - \mathbf{x}^*\|$ in order to choose the learning rate. Smoothness of f translates this to the difference of the function values.

Here (i) $\delta_t := \frac{\eta_t}{4R}$; (ii) $P_T := \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t$, which is lower bounded as $P_T \geq \frac{\xi}{4} \ln \left(\frac{T+a-1}{a} \right)$; and (iii) \mathbf{z}_T is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability δ_t/P_T .

Note that [Theorem 2](#) gives a convergence rate of $\mathcal{O}\left(\frac{1}{\log T}\right)$. We prove it in [Appendix B.7](#).

Theorem 3 (Smooth and strongly convex case with a decaying learning rate). *Let $f^{(r)}(\mathbf{x})$ be L -smooth and μ -strongly convex. Let $QComp_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a compression operator whose compression coefficient is equal to $\gamma \in (0, 1]$. Let $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ be generated according to [Algorithm 1](#) with $QComp_k$, for step sizes $\eta_t = 8/\mu(a+t)$ with $\text{gap}(\mathcal{I}_T) \leq H$, where $a > 1$ is such that we have $a > \max\{4H/\gamma, 32\kappa, H\}$, $\kappa = L/\mu$. Then the following holds*

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f^* \leq \frac{La^3}{4S_T} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{8LT(T+2a)}{\mu^2 S_T} A + \frac{128LT}{\mu^3 S_T} B.$$

Here (i) $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$, $B = 4 \left(\left(\frac{3\mu}{2} + 3L \right) \frac{CG^2H^2}{\gamma^2} + 3L^2G^2H^2 \right)$, where $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$; (ii) $\bar{\mathbf{x}}_T := \frac{1}{S_T} \sum_{t=0}^{T-1} \left[w_t \left(\frac{1}{R} \sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)} \right) \right]$, where $w_t = (a+t)^2$; and (iii) $S_T = \sum_{t=0}^{T-1} w_t \geq \frac{T^3}{3}$.

Corollary 3. *For $a > \max\{\frac{4H}{\gamma}, 32\kappa, H\}$, $\sigma_{max} = \max_{r \in [R]} \sigma_r$, and using $\mathbb{E}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{4G^2}{\mu^2}$ from [Lemma 2](#) in [\[RSS12\]](#), we have*

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f^* \leq \mathcal{O} \left(\frac{G^2H^3}{\mu^2\gamma^3T^3} \right) + \mathcal{O} \left(\frac{\sigma_{max}^2}{\mu^2bRT} + \frac{H\sigma_{max}^2}{\mu^2bR\gamma T^2} \right) + \mathcal{O} \left(\frac{G^2H^2}{\mu^3\gamma^2T^2} \right).$$

In order to ensure that the compression does not affect the dominating terms while converging at a rate of $\mathcal{O}(1/(bRT))$, we would require $H = \mathcal{O} \left(\gamma \sqrt{T/(bR)} \right)$.

[Theorem 3](#) is proved in [Appendix B.8](#). For no compression and only local computations, i.e., for $\gamma = 1$, and under the same assumptions, we recover/generalize a few recent results from literature with similar convergence rates:

1. We recover [\[YYZ19, Theorem 1\]](#), which does local SGD for the non-convex case;
2. We generalize [\[Sti19, Theorem 2.2\]](#), which does local SGD for a strongly convex case and requires the unbiasedness assumption of gradients,³ to the distributed case.

We emphasize that unlike [\[YYZ19, Sti19\]](#), which only consider local computation, we combine quantization and sparsification with local computation, which poses several technical challenges; e.g., see proofs of [Lemma 4, 5, 6](#).

3.4 Proof Outlines

In order to prove our results, we define virtual sequences for every worker $r \in [R]$ and for all $t \geq 0$ as follows:

$$\widetilde{\mathbf{x}}_0^{(r)} := \widehat{\mathbf{x}}_0^{(r)} \quad \text{and} \quad \widetilde{\mathbf{x}}_{t+1}^{(r)} := \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}} \left(\widehat{\mathbf{x}}_t^{(r)} \right) \quad (4)$$

Here η_t can be taken to be decaying or fixed, depending on the result that we are proving. Let i_t be the set of random sampling of the mini-batches at each worker $\{i_t^{(1)}, i_t^{(2)}, \dots, i_t^{(R)}\}$. We define

1. $\mathbf{p}_t := \frac{1}{R} \sum_{r=1}^R \nabla f_{i_t^{(r)}} \left(\widehat{\mathbf{x}}_t^{(r)} \right)$, $\bar{\mathbf{p}}_t := \mathbb{E}_{i_t}[\mathbf{p}_t] = \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)} \left(\widehat{\mathbf{x}}_t^{(r)} \right)$;
2. $\widetilde{\mathbf{x}}_{t+1} := \frac{1}{R} \sum_{r=1}^R \widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t - \eta_t \mathbf{p}_t$, $\widehat{\mathbf{x}}_t := \frac{1}{R} \sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$.

³The unbiasedness of gradients at every worker can be ensured by assuming that each worker samples data points from the *entire* dataset.

3.4.1 Proof Outline of Theorem 1

Proof. Since f is L -smooth, we have from (4) (with fixed learning rate $\eta_t = \eta$) that

$$f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) \leq -\eta \langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{p}_t \rangle + \frac{\eta^2 L}{2} \|\mathbf{p}_t\|_2^2. \quad (5)$$

With some algebraic manipulations provided in Appendix B.6, for $\eta \leq 1/2L$, we arrive at

$$\begin{aligned} \frac{\eta}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|_2^2 &\leq (\mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})]) + \eta^2 L \mathbb{E} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|_2^2 + 2\eta L^2 \mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2 \\ &\quad + 2\eta L^2 \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2. \end{aligned} \quad (6)$$

Under the Assumption 2, stated in Section 3.1, we have

$$\mathbb{E} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|_2^2 \leq \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}. \quad (7)$$

To bound $\mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2$ on the RHS of (6), we first show below that $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)}$, i.e., the difference of the true and the virtual sequence is equal to the average memory; and then we can use the bound on the local memory terms from Lemma 5.

Lemma 6 (Memory). *Let $\tilde{\mathbf{x}}_t^{(r)}, m_t^{(r)}, r \in [R], t \geq 0$ be generated according to Algorithm 1 and let $\hat{\mathbf{x}}_t^{(r)}$ be as defined in (4). Let $\tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_t^{(r)}$ and $\hat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)}$. Then we have*

$$\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)},$$

i.e., the difference of the true and the virtual sequence is equal to the average memory.

A proof of Lemma 6 is provided in Appendix B.3. Since $\mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2 \leq \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|m_t^{(r)}\|_2^2$, by using Lemma 5 to bound the local memory terms $\mathbb{E} \|m_t^{(r)}\|_2^2$, we get

$$\mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2 \leq 4 \frac{\eta^2 (1 - \gamma^2)}{\gamma^2} H^2 G^2. \quad (8)$$

The last term on the RHS of (6) depicts the deviation of the local sequences $\hat{\mathbf{x}}_t^{(r)}$ from the global sequence $\tilde{\mathbf{x}}_t$ which can be bounded as shown in Lemma 7. The details are provided in Appendix B.4.

Lemma 7 (Bounded deviation of local sequences). *Let $\text{gap}(\mathcal{I}_T) \leq H$. For $\hat{\mathbf{x}}_t^{(r)}$ generated according to Algorithm 1 with a fixed learning rate η and letting $\hat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)}$, we have the following bound on the deviation of the local sequences:*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2 \leq \eta^2 G^2 H^2. \quad (9)$$

Substituting the bounds from (7)-(9) into (6) yields

$$\begin{aligned} \frac{\eta}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|_2^2 &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})] + \frac{\eta^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 8 \frac{\eta^3 (1 - \gamma^2)}{\gamma^2} L^2 G^2 H^2 \\ &\quad + 2\eta^3 L^2 G^2 H^2. \end{aligned} \quad (10)$$

Performing a telescopic sum from $t = 0$ to $T - 1$ and dividing by $\frac{\eta T}{4}$ gives

$$\frac{1}{RT} \sum_{t=0}^{T-1} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 \leq \frac{4(\mathbb{E}[f(\widetilde{\mathbf{x}}_0)] - f^*)}{\eta T} + \frac{4\eta L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 32 \frac{\eta^2(1-\gamma^2)}{\gamma^2} L^2 G^2 H^2 + 8\eta^2 L^2 G^2 H^2. \quad (11)$$

By letting $\eta = \widehat{C}/\sqrt{T}$, where \widehat{C} is a constant such that $\frac{\widehat{C}}{\sqrt{T}} \leq \frac{1}{2L}$, we arrive at bound stated in [Theorem 1](#). \square

3.4.2 Proof Outline of [Theorem 2](#)

Proof. Observe that [\(6\)](#) holds irrespective of the learning rate schedule, as long as learning rate is at most $1/2L$; see [Appendix B.7](#) for details. Here $\eta_t \leq \frac{1}{2L}$ follows from our assumption that $a \geq 2\xi L$. Substituting a decaying learning rate η_t (such that $\eta_t \leq 1/2L$ holds for every $t \geq 0$) in [\(6\)](#) gives

$$\begin{aligned} \frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 &\leq (\mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})]) + \eta_t^2 L \mathbb{E} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|_2^2 + 2\eta_t L^2 \mathbb{E} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|_2^2 \\ &\quad + 2\eta_t L^2 \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2. \end{aligned} \quad (12)$$

We have already bounded $\mathbb{E} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|_2^2 \leq \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$ in [\(7\)](#). Note that [Lemma 6](#) holds irrespective of the learning rate schedule, and together with [Lemma 4](#), we can show that

$$\mathbb{E} \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 \leq C \frac{4\eta_t^2}{\gamma^2} G^2 H^2. \quad (13)$$

The last term on the RHS of [\(12\)](#) is the deviation of local sequences and we bound it in [Lemma 8](#) for decaying learning rates. The details are provided in [Appendix B.5](#).

Lemma 8 (Contracting deviation of local sequences). *Let $\text{gap}(\mathcal{I}_T) \leq H$. By running [Algorithm 1](#) with a decaying learning rate η_t , we have*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq 4\eta_t^2 G^2 H^2. \quad (14)$$

Observe that for the case of fixed learning rate, we can trivially bound $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2$ by simply putting $\eta_t = \eta$ in [\(14\)](#). However, in [\(9\)](#), we can get a slightly better bound (without the factor of 4) by directly working with a fixed learning rate. Using these bounds in [\(12\)](#) gives

$$\frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + \frac{8\eta_t^3}{\gamma^2} CL^2 G^2 H^2 + 8\eta_t^3 L^2 G^2 H^2.$$

Let $\delta_t := \frac{\eta_t}{4R}$ and $P_T := \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t$. Performing a telescopic sum from $t = 0$ to $T - 1$ and dividing by P_T gives

$$\begin{aligned} \frac{1}{P_T} \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 &\leq \frac{\mathbb{E}f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2}{bR^2(a-1)} \frac{\sum_{r=1}^R \sigma_r^2}{P_T} \\ &\quad + \left(\frac{8C}{\gamma^2} + 8 \right) L^2 G^2 H^2 \frac{\xi^3}{2P_T(a-1)^2} \end{aligned} \quad (15)$$

In (15), we used the following bounds, which are shown in [Appendix B.7](#): $P_T \geq \frac{\xi}{4} \ln \left(\frac{T+a-1}{a} \right)$, $\sum_{t=0}^{T-1} \eta_t^2 \leq \frac{\xi^2}{a-1}$, and $\sum_{t=0}^{T-1} \eta_t^3 \leq \frac{\xi^3}{2(a-1)^2}$. This completes the proof of [Theorem 2](#). \square

3.4.3 Proof Outline of [Theorem 3](#)

Proof. Using the definition of virtual sequences (4) that, we have

$$\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 = \|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|_2^2 + \eta_t^2 \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|_2^2 - 2\eta_t \langle \tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t, \mathbf{p}_t - \bar{\mathbf{p}}_t \rangle. \quad (16)$$

Note that $\eta_t \leq 1/4L$, which follows from the assumption that $a > \frac{32L}{\mu}$. Now, using μ -strong convexity and L -smoothness of f , together with some algebraic manipulations provided in [Appendix B.8](#), by letting $e_t = \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - f^*$, we arrive at

$$\begin{aligned} \mathbb{E}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|_2^2 - \frac{\eta_t\mu}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) \mathbb{E}\|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2 \\ &\quad + \frac{3\eta_t L}{R} \sum_{r=1}^R \mathbb{E}\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}. \end{aligned} \quad (17)$$

Note that the bounds in (13) and (14) hold irrespective of whether the function is convex or not. So, we can use them here as well in (17), which gives

$$\begin{aligned} \mathbb{E}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|_2^2 - \frac{\mu\eta_t}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) C \frac{4\eta_t^2}{\gamma^2} G^2 H^2 \\ &\quad + (3\eta_t L) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}. \end{aligned} \quad (18)$$

Employing a slightly modified result than [[SCJ18](#), Lemma 3.3] with $a_t = \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|_2^2$, $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$ and $B = 4 \left(\left(\frac{3\mu}{2} + 3L \right) \frac{CG^2H^2}{\gamma^2} + 3L^2G^2H^2 \right)$, we have

$$a_{t+1} \leq \left(1 - \frac{\mu\eta_t}{2}\right) a_t - \frac{\mu\eta_t}{2L} e_t + \eta_t^2 A + \eta_t^3 B. \quad (19)$$

For $\eta_t = \frac{8}{\mu(a+t)}$ and $w_t = (a+t)^2$, $S_T = \sum_{t=0}^{T-1} \geq \frac{T^3}{3}$, we have

$$\frac{\mu}{2LS_T} \sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu a^3}{8S_T} a_0 + \frac{4T(T+2a)}{\mu S_T} A + \frac{64T}{\mu^2 S_T} B. \quad (20)$$

From convexity, we can finally write

$$\mathbb{E}f(\bar{\mathbf{x}}_T) - f^* \leq \frac{La^3}{4S_T} a_0 + \frac{8LT(T+2a)}{\mu^2 S_T} A + \frac{128LT}{\mu^3 S_T} B. \quad (21)$$

Where $\bar{\mathbf{x}}_T := \frac{1}{S_T} \sum_{t=0}^{T-1} \left[w_t \left(\frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)} \right) \right] = \frac{1}{S_T} \sum_{t=0}^{T-1} w_t \hat{\mathbf{x}}_t$. This completes the proof of [Theorem 3](#). \square

4 Distributed Asynchronous Operation

We propose and analyze a particular form of asynchronous operation, where the workers synchronize with the master at arbitrary times decided locally or by master picking a subset of nodes as in federated learning [[Kon17](#), [MMR⁺17](#)]. However, the local iterates evolve at the same

rate, i.e., each worker takes the same number of steps per unit time according to a global clock. The asynchrony is therefore that updates occur after different number of local iterations but the local iterations are in synchrony with respect to the global clock. This is different from asynchronous algorithms studied for stragglers [WYL⁺18, RRWN11], where only one gradient step is taken but occurs at different times due to delays.

In this asynchronous setting, $\mathcal{I}_T^{(r)}$'s may be different for different workers. However, we assume that $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$, which means that there is a uniform bound on the maximum delay in each worker's update times. The algorithmic difference from [Algorithm 1](#) is that, in this case, a subset of workers (including a single worker) can send their updates to the master at their synchronization time steps; master aggregates them, updates the global parameter vector, and sends that only to those workers. Our algorithm is summarized in [Algorithm 2](#)

Algorithm 2 Qsparse-local-SGD with asynchronous updates

```

1: Initialize  $\mathbf{x}_0 = \bar{\mathbf{x}}_0 = \mathbf{x}_0^{(r)} = \hat{\mathbf{x}}_0^{(r)} = m_0^{(r)} = \mathbf{0}, \forall r \in [R]$ . Suppose  $\eta_t$  follows a certain learning rate
   schedule.
2: for  $t = 0$  to  $T - 1$  do
3:   On Workers:
4:   for  $r = 1$  to  $R$  do
5:      $\hat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} \leftarrow \hat{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}(\hat{\mathbf{x}}_t^{(r)})$ ;  $i_t^{(r)}$  is a mini-batch of size  $b$  uniformly in  $\mathcal{D}_r$ 
6:     if  $t + 1 \notin \mathcal{I}_T^{(r)}$  then
7:        $\mathbf{x}_{t+1}^{(r)} \leftarrow \mathbf{x}_t^{(r)}, m_{t+1}^{(r)} \leftarrow m_t^{(r)}$  and  $\hat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \hat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$ 
8:     else
9:        $g_t^{(r)} \leftarrow Q \text{Comp}_k(m_t^{(r)} + \mathbf{x}_t^{(r)} - \hat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)})$  and send  $g_t^{(r)}$  to the master
10:       $m_{t+1}^{(r)} \leftarrow m_t^{(r)} + \mathbf{x}_t^{(r)} - \hat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} - g_t^{(r)}$ 
11:      Receive  $\bar{\mathbf{x}}_{t+1}$  from the master and set  $\mathbf{x}_{t+1}^{(r)} \leftarrow \bar{\mathbf{x}}_{t+1}$  and  $\hat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \bar{\mathbf{x}}_{t+1}$ 
12:    end if
13:  end for
14:  At Master:
15:  if  $t + 1 \notin \mathcal{I}_T^{(r)}$  for all  $r \in [R]$  then
16:     $\bar{\mathbf{x}}_{t+1} \leftarrow \bar{\mathbf{x}}_t$ 
17:  else
18:    Let  $\mathcal{S} \subseteq [R]$  be the set of all workers  $r$  such that master receives  $g_t^{(r)}$  from  $r$ 
19:    Compute  $\bar{\mathbf{x}}_{t+1} \leftarrow \bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r \in \mathcal{S}} g_t^{(r)}$  and broadcast  $\bar{\mathbf{x}}_{t+1}$  to all the workers in  $\mathcal{S}$ 
20:  end if
21: end for

```

4.1 Main Results

In this section we present our main convergence results with asynchronous updates, obtained by running [Algorithm 2](#) for smooth objectives, both non-convex and strongly convex. Under the same assumptions as in the synchronous setting of [Section 3.1](#), the following theorems hold even if [Algorithm 2](#) is run with an arbitrary compression operators (including our composed operators from [Section 2.3](#)), whose compression coefficient is equal to γ .

Theorem 4 (Smooth (non-convex) case with fixed learning rate). *Under the same conditions as in [Theorem 1](#) with $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$, if $\{\hat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to [Algorithm 2](#), the*

following holds.

$$\begin{aligned} \mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 &\leq \left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{\widehat{C}} + \widehat{C}L \left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \right) \right) \frac{4}{\sqrt{T}} \\ &\quad + 8 \left(12 \frac{(1-\gamma^2)}{\gamma^2} + (2 + 8C_1H^2) \right) \frac{\widehat{C}^2 L^2 G^2 H^2}{T}. \end{aligned}$$

Here (i) $C_1 = (\frac{8}{\gamma^2} - 6)(4 - 2\gamma)$; (ii) \mathbf{z}_T is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability $1/RT$; and (iii) \widehat{C} is a constant such that $\frac{\widehat{C}}{\sqrt{T}} \leq \frac{1}{2L}$.

Corollary 4. Let $\mathbb{E}[f(\mathbf{x}_0)] - f^* \leq J^2$, where $J < \infty$ is a constant, $\sigma_{max} = \max_{r \in [R]} \sigma_r$, and $\widehat{C}^2 = bR(\mathbb{E}[f(\mathbf{x}_0)] - f^*)/\sigma_{max}^2 L$. We can get a simplified expression below

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \mathcal{O} \left(\frac{J\sigma_{max}}{\sqrt{bRT}} \right) + \mathcal{O} \left(\frac{J^2 b R G^2}{\sigma_{max}^2 \gamma^2 T} (H^2 + H^4) \right).$$

In order to ensure that the compression does not affect the dominating terms while converging at a rate of $\mathcal{O}(1/\sqrt{bRT})$, we would require $H = \mathcal{O}(\sqrt{\gamma T^{1/8}}/(bR)^{3/8})$.

Theorem 4 provides non asymptotic guarantees where we also observe that the compression comes for "free". The corresponding asymptotic result is given below.

Theorem 5 (Smooth (non-convex) case with decaying learning rate). Under the same conditions as in **Theorem 2** with $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to **Algorithm 2**, the following holds.

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2}{(a-1)P_T} \left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \right) + \left(16 + \frac{24C}{\gamma^2} + 200C'H^2 \right) \frac{\xi^3 L^2 G^2 H^2}{2(a-1)^2 P_T}$$

Here (i) $\delta_t := \frac{\eta_t}{4R}$ and $P_T := \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t$, which is lower bounded as $P_T \geq \frac{\xi}{4} \ln \left(\frac{T+a-1}{a} \right)$; (ii) $C' = (4 - 2\gamma)(1 + \frac{C}{\gamma^2})$; and (iii) \mathbf{z}_T is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability δ_t/P_T .

Theorem 6 (Smooth and strongly convex case with decaying learning rate). Under the same conditions as in **Theorem 3** with $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to **Algorithm 2**, the following holds.

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f^* \leq \frac{La^3}{4S_T} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{8LT(T+2a)}{\mu^2 S_T} A + \frac{128LT}{\mu^3 S_T} D$$

Here (i) $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$, $C_1 = 192(4 - 2\gamma) \left(1 + \frac{C}{\gamma^2} \right)$, $C_2 = 8(4 - 2\gamma)(1 + \frac{C}{\gamma^2})$; (ii) $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$, $D = \left(\frac{3\mu}{2} + 3L \right) \left(\frac{12CG^2H^2}{\gamma^2} + C_1\eta_t^2 H^4 G^2 \right) + 24(1 + C_2 H^2) L G^2 H^2$; and (iii) $\bar{\mathbf{x}}_T, S_T$ are as defined in **Theorem 3**.

Corollary 5. Under the same conditions as in **Theorem 3** with $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$, $a > \max\{\frac{4H}{\gamma}, 32\kappa, H\}$, $\sigma_{max} = \max_{r \in [R]} \sigma_r$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to **Algorithm 2**, the following holds:

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] - f^* \leq \mathcal{O} \left(\frac{G^2 H^3}{\mu^2 \gamma^3 T^3} \right) + \mathcal{O} \left(\frac{\sigma_{max}^2}{\mu^2 bRT} + \frac{H\sigma_{max}^2}{\mu^2 bR\gamma T^2} \right) + \mathcal{O} \left(\frac{G^2}{\mu^3 \gamma^2 T^2} (H^2 + H^4) \right),$$

where $\bar{\mathbf{x}}_T, S_T$ are as defined in **Theorem 3**. In order to ensure that the compression does not affect the dominating terms while converging at a rate of $\mathcal{O}(1/(bRT))$, we would require $H = \mathcal{O}(\sqrt{\gamma(T/(bR))}^{1/4})$.

4.2 Proof Outlines

Our proofs of these results follow the same outlines of the corresponding proofs in the synchronous setting, but some technical details change significantly, which arise because, in our asynchronous setting, workers are allowed to update the global parameter vector in between two consecutive synchronization time steps of other workers. Specifically, in the asynchronous setting, we have to bound the deviation of local sequences $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_t^{(r)}\|_2^2$ and the difference between the virtual and true sequences $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2$, both with a fixed learning rate as well as with decaying learning rate. We show these below in [Lemma 9-10](#) and [Lemma 11-12](#).

Lemma 9 (Contracting local sequence deviation). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. For $\hat{\mathbf{x}}_t^{(r)}$ generated according to [Algorithm 2](#) with decaying learning rate η_t and letting $\hat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)}$, we have the following bound on the deviation of the local sequences:*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2 \leq 8(1 + C'' H^2) \eta_t^2 G^2 H^2,$$

where $C'' = 8(4 - 2\gamma)(1 + \frac{C}{\gamma^2})$ and C is a constant satisfying $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$.

Lemma 10 (Bounded local sequence deviation). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. By running [Algorithm 2](#) with fixed learning rate η , we have*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2 \leq (2 + H^2 C') \eta^2 G^2 H^2,$$

where $C' = (\frac{16}{\gamma^2} - 12)(4 - 2\gamma)$.

We prove these above two lemmas in [Appendix C.1](#) and [Appendix C.2](#), respectively. Note that the bound in [Lemma 9](#) is $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|_2^2 \leq \mathcal{O}(\eta_t^2 G^2 (H^2 + H^4/\gamma^2))$, which is weaker than the corresponding bound $\mathcal{O}(\eta_t^2 G^2 H^2)$ for the synchronous setting in [Lemma 8](#). See [Lemma 10](#) and [Lemma 7](#) for a similar comparison for the case of fixed learning rate.

Now we bound $\mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|_2^2$. Fix a time t and consider any worker $r \in [R]$. Let $t_r \in \mathcal{I}_T^{(r)}$ denote the last synchronization step until time t for the r 'th worker. Define $t'_0 := \min_{r \in [R]} t_r$. We want to bound $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2$. Note that in the synchronous case, we have shown in [Lemma 6](#) that $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)}$. This does not hold in the asynchronous setting, which makes upper-bounding $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2$ a bit more involved. By definition $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R (\hat{\mathbf{x}}_t^{(r)} - \tilde{\mathbf{x}}_t^{(r)})$. By the definition of virtual sequences and the update rule for $\hat{\mathbf{x}}_t^{(r)}$, we also have $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R (\hat{\mathbf{x}}_{t_r}^{(r)} - \tilde{\mathbf{x}}_{t_r}^{(r)})$. This can be written as

$$\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \left[\frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0} \right] + \left[\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t \right] + \left[\bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_{t_r}^{(r)} \right]. \quad (22)$$

In (22), the third term on the RHS is equal to the average memory as shown in (96) in [Appendix C.3](#), and unlike [Lemma 6](#) in the synchronous setting, which states that $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)}$, does not hold here. However, we can show that $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t$ is equal to the sum of $\frac{1}{R} \sum_{r=1}^R m_t^{(r)}$ and an additional term, which leads to potentially a weaker bound $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2 \leq \mathcal{O}(\eta_t^2/\gamma^2 G^2 (H^2 + H^4))$, proved in [Lemma 11-12](#) in [Appendix C.3](#) and [Appendix C.4](#), in comparison to $\mathcal{O}(\eta_t^2/\gamma^2 G^2 H^2)$ for the synchronous setting.

Lemma 11 (Contracting distance between virtual and true sequence). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. If we run [Algorithm 2](#) with a decaying learning rate η_t , then we have the following bound on the difference between the true and virtual sequences:*

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 \leq C'\eta_t^2 H^4 G^2 + 12C\frac{\eta_t^2}{\gamma^2} G^2 H^2,$$

where $C' = 192(4 - 2\gamma) \left(1 + \frac{C}{\gamma^2}\right)$ and C is a constant satisfying $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$.

Lemma 12 (Bounded distance between virtual and true sequence). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. If we run [Algorithm 2](#) with a fixed learning rate η , we have*

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 \leq 6C'\eta^2 H^4 G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2} G^2 H^2,$$

where $C' = (4 - 2\gamma) \left(\frac{8}{\gamma^2} - 6\right)$.

Summary of our results. Now we give a brief summary of our convergence results in the synchronous as well as asynchronous settings.

1. In the synchronous setting, *Qsparse-local-SGD* asymptotically converges as fast as distributed vanilla SGD for $H = \mathcal{O}(\gamma T^{1/4}/(bR)^{3/4})$ in the smooth and non-convex case and for $H = \mathcal{O}(\gamma\sqrt{T/(bR)})$ in the strongly convex case.
2. In the asynchronous setting, *Qsparse-local-SGD* asymptotically converges as fast as distributed vanilla SGD for $H = \mathcal{O}(\sqrt{\gamma}T^{1/8}/(bR)^{3/8})$ in the smooth and non-convex case and for $H = \mathcal{O}(\sqrt{\gamma}(T/(bR))^{1/4})$ in the strongly convex case.

Therefore, our algorithm provides a lot of flexibility in terms of different ways of mitigating the communication bottleneck. For example, by increasing the batch size on each node, or by increasing the maximum synchronization period H up to allowable limits. Furthermore, one could also choose to opt for different values of k for the Top_k sparsifier, as well as adjust the configurations of the quantizer. We present numerics in [Section 5](#) demonstrating significant savings in the number of bits exchanged over the state-of-the-art.

5 Experimental Results

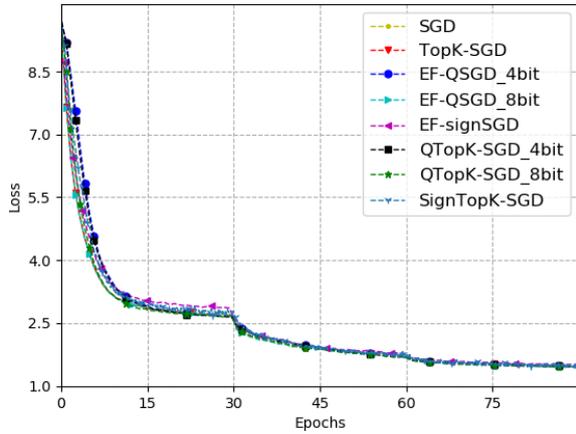
In this section we give extensive experimental results for validating our theoretical findings.

5.1 Non-Convex Objective

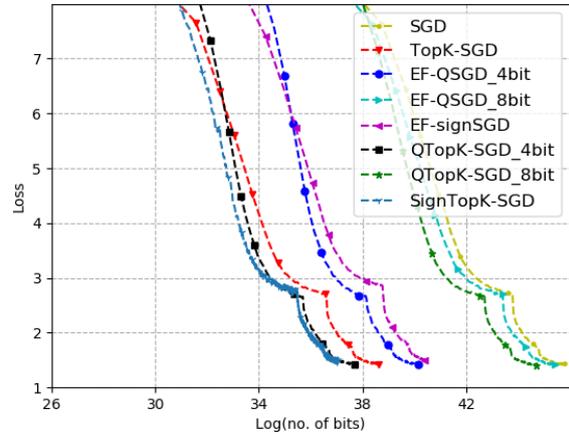
5.1.1 Experiment Setup

We train ResNet-50 [[HZRS16](#)] (which has $d = 25,610,216$ parameters) on ImageNet dataset, using 8 NVIDIA Tesla V100 GPUs. We use a learning rate schedule consisting of 5 epochs of linear warmup, followed by a piecewise decay of 0.1 at epochs 30, 60 and 80, with a batch size of 256 per GPU. For the purpose of experiments, we focus on SGD with momentum of 0.9, applied on the local iterations of the workers. We build our compression scheme into the Horovod framework [[SB18](#)]. We use SignTop_k as defined in [Lemma 3](#) and QTop_k as defined in [Lemma 1](#) (which has an operating regime $\beta_{k,s} < 1$), where Q is from [[AGL+17](#)], as our

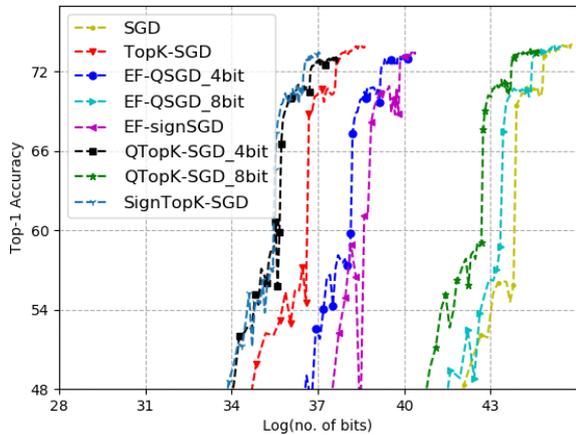
composed operators.⁴ In Top_k , we only update $k_t = \min(d_t, 1000)$ elements per step for each tensor t , where d_t is the number of elements in the tensor. For ResNet-50 architecture, this amounts to updating a total of $k = 99,400$ elements per step.



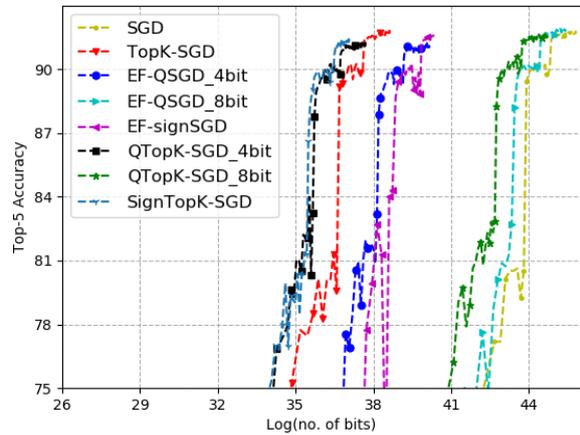
(a) Training loss vs epochs



(b) Training loss vs \log_2 of communication budget



(c) top-1 accuracy [LHS15] for schemes in Figure 1a



(d) top-5 accuracy [LHS15] for schemes in Figure 1a

Figure 1 Figure 1a-1d demonstrate the gains in performance achieved by our Q sparse operators in the non-convex setting.

5.1.2 Results

From Figure 1a, we observe that quantization and sparsification, both individually and combined, when error compensation is enabled through accumulating errors, has almost no penalty in terms of convergence rate, with respect to vanilla SGD. We observe that both $QTop_k$ -SGD, which employs a 4 bit quantizer and the Top_k sparsifier, as well as $SignTop_k$ -SGD, which employs the 1 bit sign quantizer and the Top_k sparsifier, demonstrate superior performance over other

⁴Even though the “scaled” $QTop_k$ from Lemma 2 (with a scaling factor of $(1 + \beta_{k,s})$) works with all values of $\beta_{k,s}$, and also does better than the “unscaled” $QTop_k$ from Lemma 1 even when $\beta_{k,s} < 1$ (see Remark 2), we report our experimental results in the non-convex setting only with unscaled $QTop_k$. We give some plots for a comparison on both these operators in Appendix D and observe that our algorithm with the unscaled operator gives at least as good performance as it gives with the scaled operator. We can attribute this to the fact that scaling the composed operator is a sufficient condition to obtain better convergence results, which does not necessarily mean that in practice also it does better.

schemes, both in terms of the required number of communicated bits for achieving certain target loss as well as test accuracy. This is because, in $QTop_k$, the Q operator from [AGL⁺17] further induces sparsity, which results in fewer than k coordinates being transmitted, and in $SignTop_k$, we send only 1 bit for each Top_k coordinate.

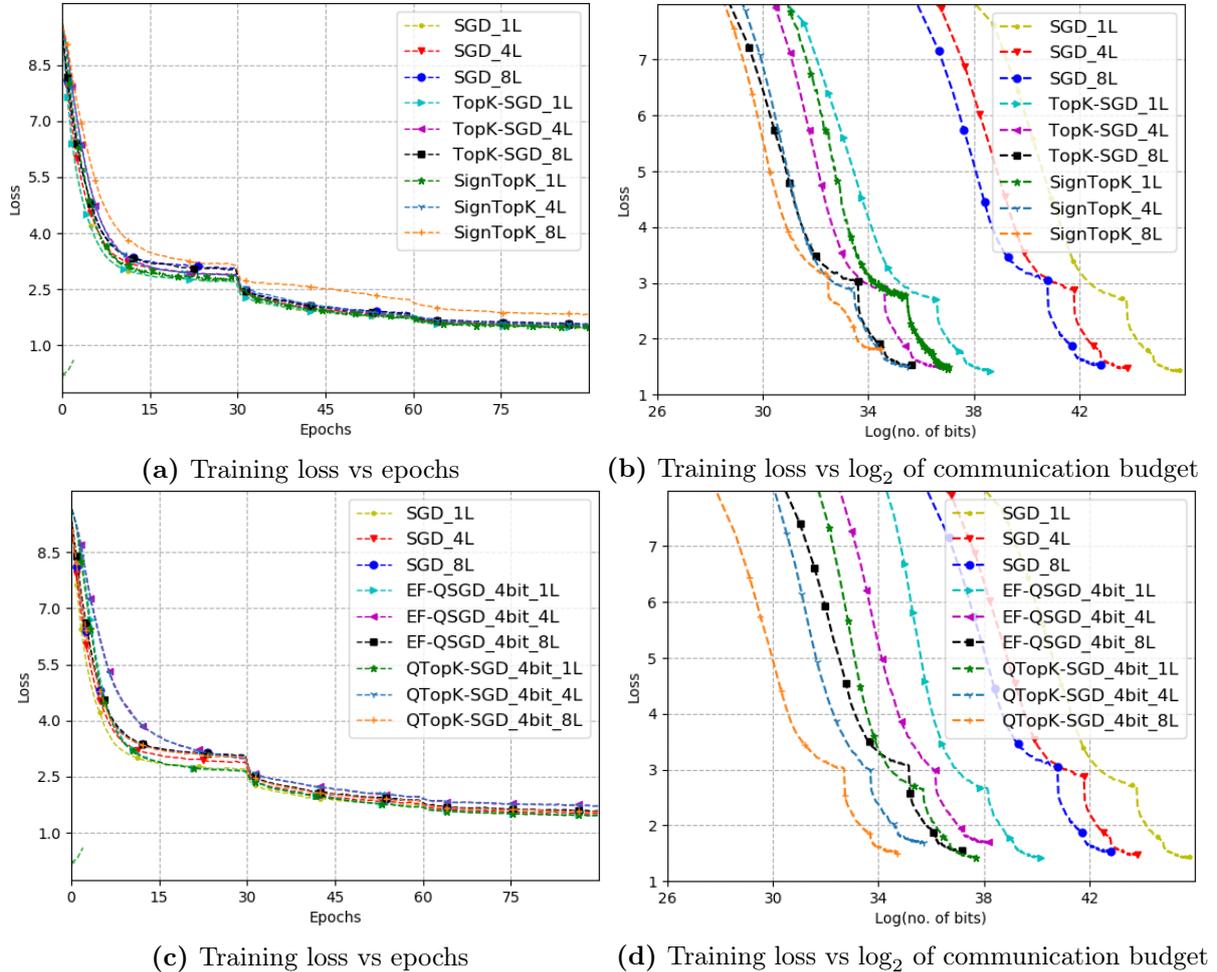
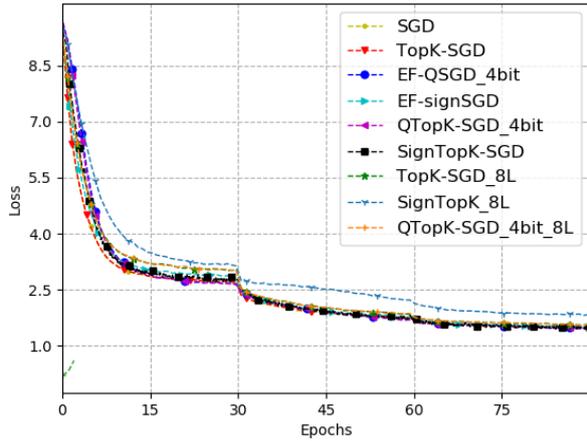


Figure 2 Figure 2a-2b demonstrate the effect of incorporating local iterations and compare these effects across vanilla SGD, the sparsifier Top_k , as well as its composition with the $Sign$ operator. Similar comparisons are also made between vanilla SGD, the quantizer QSGD with error accumulation, as well as its composition with the Top_k sparsifier. $SignTopK_hL$, for $h = 1, 4, 8$, corresponds to running Algorithm 1 with $SignTopK$ being the composed operator with a synchronization period of at most h .

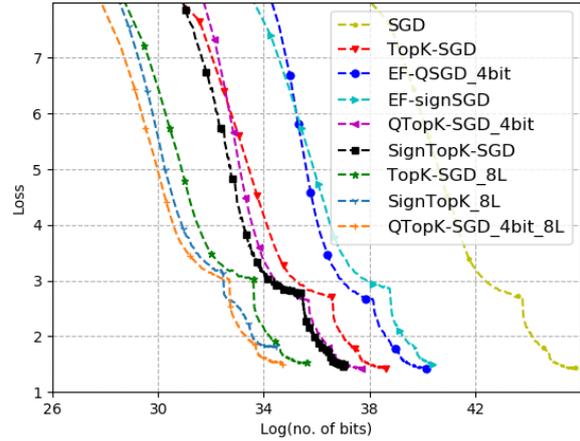
In Figure 2a-2d, we show how the performance of different methods (used in Figure 1a-1d) change when we incorporate local iterations on top of them. Observe that the incorporation of local iterations in Figure 2a and 2c has very little impact on the convergence rates, as compared to vanilla SGD with the corresponding number of local iterations. Furthermore, this provides an added advantage over the $Qsparse$ operator, in terms of savings in communicated bits for achieving target loss as seen in Figure 2b and 2d, by a factor of 6 to 8 times on average.

Figure 3b, Figure 3c, and Figure 3d show the training loss, top-1, and top-5 convergence rates⁵ respectively, with respect to the total number of bits of communication used. We observe that $Qsparse-local-SGD$ combines the bit savings of either the deterministic sign based

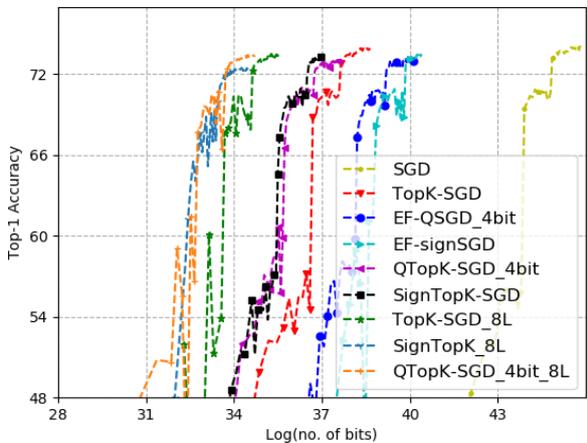
⁵Here top- i refers to the accuracy of the top i predictions by the model from the list of possible classes, see [LHS15].



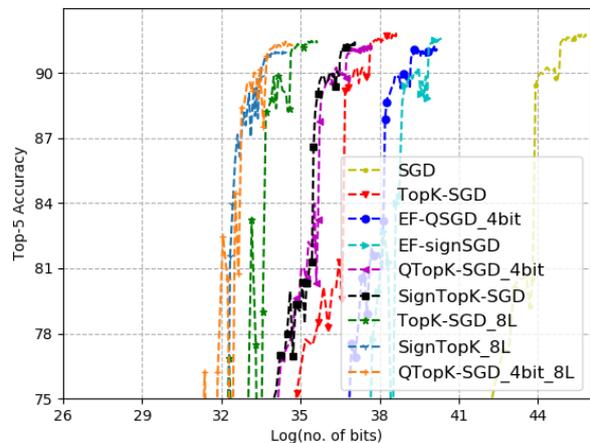
(a) Training loss vs against epochs



(b) Training loss vs \log_2 of communication budget



(c) top-1 accuracy [LHS15] for schemes in Figure 3a



(d) top-5 accuracy [LHS15] for schemes in Figure 3a

Figure 3 Figure 3a-3d demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRSJ19], TopK-SGD [SCJ18, AHJ⁺18] and local SGD [Sti19, YYZ19] in the non-convex setting.

operator or the stochastic quantizer (QSGD), and aggressive sparsifier along with infrequent communication, thereby, outperforming the cases where these techniques are individually used. In particular, the required number of bits to achieve the same loss or top-1 accuracy in the case of *Qsparse-local-SGD* is around 1/16 in comparison with *Top_k-SGD* and over 1000 \times less than vanilla SGD. This also verifies that error compensation through memory can be used to mitigate not only the missing components from updates in previous synchronization rounds, but also explicit quantization error.

5.2 Convex Objective

The experiments in Figure 4-6 are in a synchronous distributed setting with 15 worker nodes, each processing a mini-batch size of 8 samples per iteration using the *MNIST* [LBBH98] handwritten digits dataset. The corresponding experiments for the asynchronous operation (as in Algorithm 2) are shown in Figure 7.

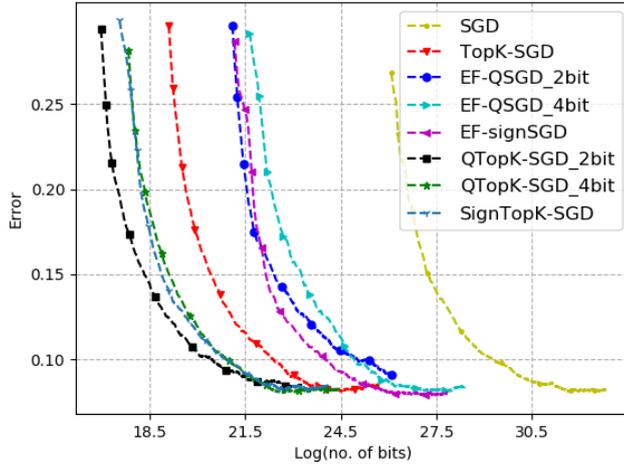
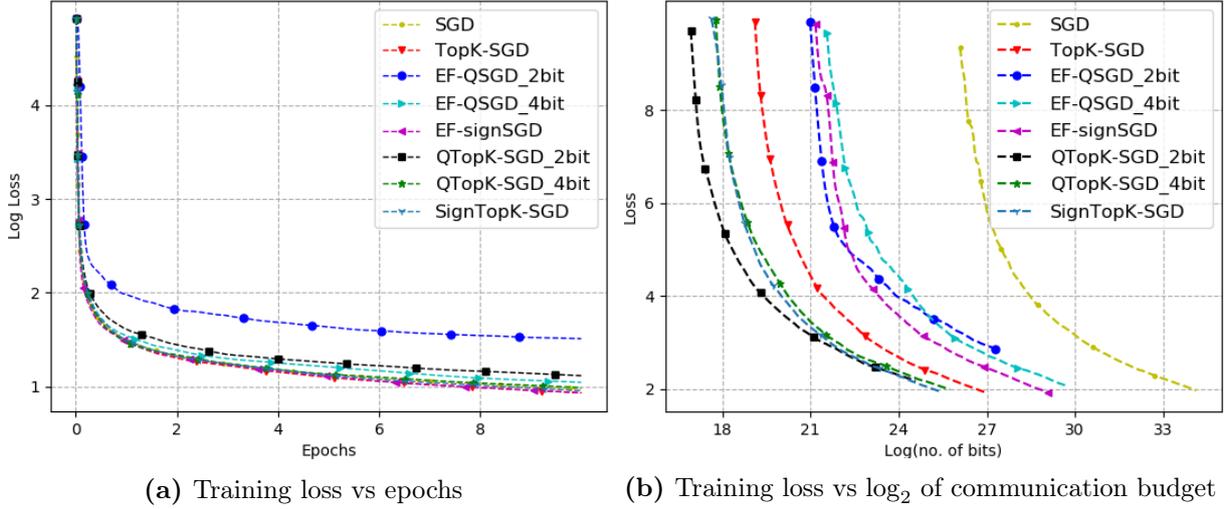


Figure 4 Figure 4a-4c demonstrate the gains in performance achieved by our Q sparse operators in the convex setting.

5.2.1 Model Architecture

Define the softmax function as

$$h_{\mathbf{x},z}(a^{(i)}) = \frac{\exp(\mathbf{x}_j^T a^{(i)} + z^{(i)})}{\sum_{l=1}^L \exp(\mathbf{x}_l^T a^{(i)} + z^{(l)})}.$$

Our experiments are all for softmax regression with a standard ℓ_2 regularizer. The cost function is

$$-\frac{1}{n} \left(\sum_{i=1}^n \sum_{j=1}^L \mathbb{1}\{b^{(i)} = j\} \log h_{\mathbf{x},z}(a^{(i)}) \right) + \frac{\lambda}{2} \|\mathbf{x}\|^2$$

where $a^{(i)} \in \mathbb{R}^d$, $b^{(i)} \in [L]$ are the data points, which can belong to one of the L classes, and $\mathbf{x}_j \in \mathbb{R}^d$ for every $j \in [L]$, are columns of the parameter structured as follows

$$\mathbf{x} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_L], \quad \mathbf{x}_j \in \mathbb{R}^d, \quad \forall j \in [L],$$

and $z^{(i)}$ for every $i \in [L]$ are the biases to be learnt corresponding to every class. We set λ to be $1/n$.

5.2.2 Parameter Selection and Learning Rates

We use the deterministic operator as in [Lemma 3](#) and the stochastic operator *QSGD* denoted by Q , as defined in [\[AGL⁺17\]](#), as our quantizers and *Top_k* with error compensation as the sparsifier. The schemes with which we compare our composed operators *QTop_k* [Lemma 2](#), and *SignTop_k* [Lemma 3](#), are EF-QSGD [\[WHHZ18\]](#), EF-SIGNSGD [\[KRSJ19\]](#), TopK-SGD [\[SCJ18, AHJ⁺18\]](#), and local SGD [\[Sti19\]](#). The learning rate used for training is of the form $\frac{c}{\lambda(a+t)}$, where (i) λ is the regularization parameter; (ii) c is set with a careful hyperparameter sweep; (iii) $w_t = (a+t)^2$ as in [Theorem 3](#), where a is set as $\frac{dH}{k}$ with d being the dimension of the gradient vector (7850 for *MNIST*); (iv) $k = 40$ is the sparsity; (v) H is the synchronization period; (vi) t is the iteration index; (vii) $b = 8$ is the batch size; and (viii) $R = 15$ is the number of workers.

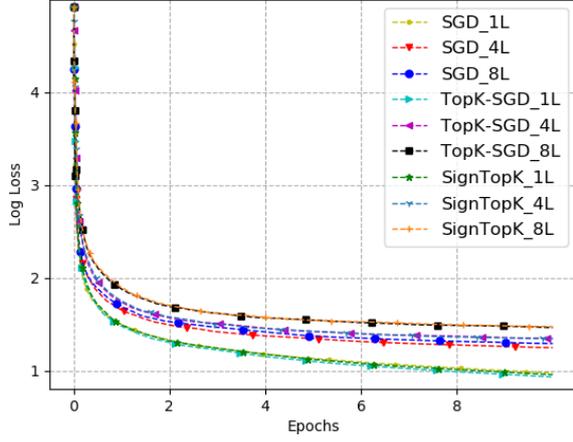
5.2.3 Results

In [Figure 4a](#), we observe that the composition of a quantizer with a sparsifier has very little effect on the rate of convergence as compared to when the techniques are used individually. Observe that the algorithm run with the 2 bit *QSGD* is slower than the 4 bit quantizer, both with or without sparsification, which can be attributed to the reduction in the compression coefficient γ in going from 4 to 2 bits; see [Theorem 3](#). From [Figure 4b](#) and [4c](#), we see that our composed operators achieve gains in communicated bits by a factor of 6-8 times over the state-of-the-art.

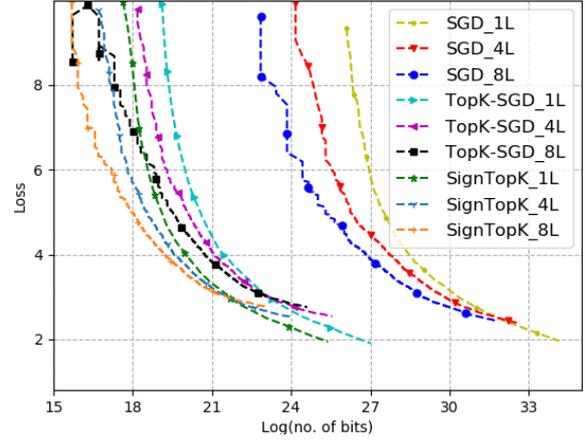
[Figure 5a](#), demonstrates the effect of incorporating local iterations together with *Qsparse* operators, and we see that the rate of convergence is not significantly affected as we go from 1 to 8 local iterations. Furthermore, observe that for a fixed number of local iterations, the *Qsparse* operator maintains the same rates as vanilla *SGD* or *Top_k-SGD*. In doing so, it is able to achieve gains in communicated bits as seen in [Figure 5b](#), simply by communicating infrequently with the master. On comparing [Figure 5c](#) and [5e](#), we observe that the *QTop_k* operator is more sensitive to the increase in local computations for coarser quantizers (smaller values of s , in this case $s = 2^{\#-bits} - 1 = 3$). This can be verified from [Figure 5e](#) which uses a 4 bit quantizer (which implies $s = 15$ instead), and the corresponding effect of local iterations on the convergence rate is less prominent. We make comparisons between vanilla *SGD*, *QSGD* with error accumulation (EF-QSGD) and our *QTop_k* operator in [Figure 5c](#) and [5d](#), for which we do not observe much difference in performance between a finer and coarser quantizer, even though the convergence rates with respect to iterations are affected. This can be attributed to the precision of the quantizer itself.

In [Figure 6a](#) and [Figure 6b](#), we compare the convergence of our proposed scheme in [Algorithm 1](#) with *QTop_k* and *SignTop_k* being the composed operators, with vanilla *SGD* (32 bit floating point), EF-QSGD, EF-SIGNSGD [\[KRSJ19\]](#), and TopK-SGD [\[SCJ18, AHJ⁺18\]](#). Both figures follow a similar trend, where we observe *QTop_k*, *SignTop_k* and TopK-SGD to be converging at the same rate as that of vanilla *SGD*, which is similar to the observations in [\[SCJ18\]](#). This implies that the composition of quantization with sparsification does not affect the convergence while achieving improved communication efficiency, as can be seen in [Figure 6c](#) and [Figure 6b](#). [Figure 6c](#) shows that for test error approximately 0.1, *Qsparse-local-SGD* combines the benefits of the composed operator *SignTop_k* or *QTop_k*, with local computations, and needs 10-15 times less bits than TopK-SGD and 1000 \times less bits than vanilla *SGD*.

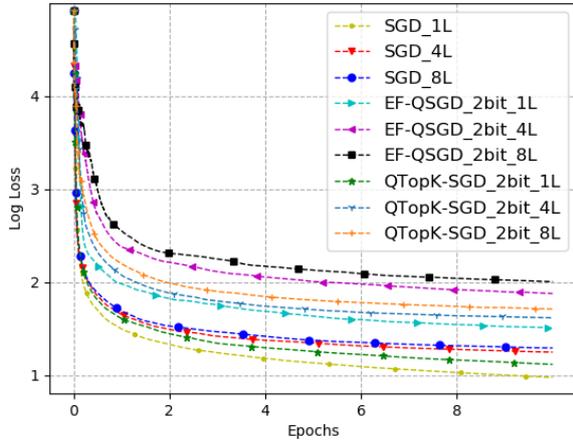
We observe similar trends in [Figure 7a-7b](#) for our asynchronous operation, where workers synchronize with the master at arbitrary time intervals as per [Algorithm 2](#). Specifically, in our experiments, for each $r \in [R]$, the time interval for the r th worker is decided uniformly at



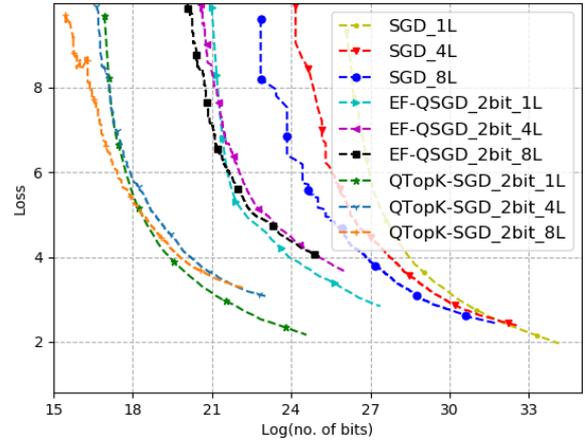
(a) Training loss vs epochs



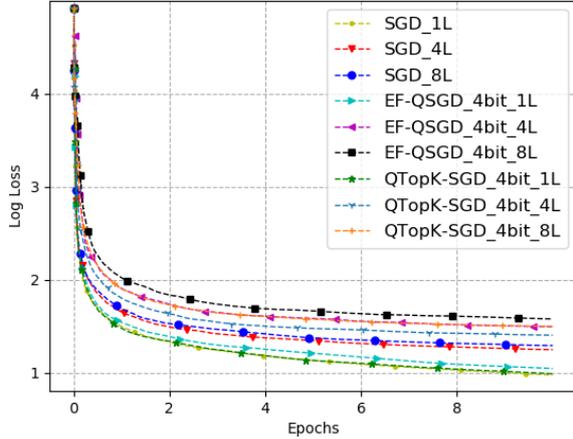
(b) Training loss vs \log_2 of communication budget



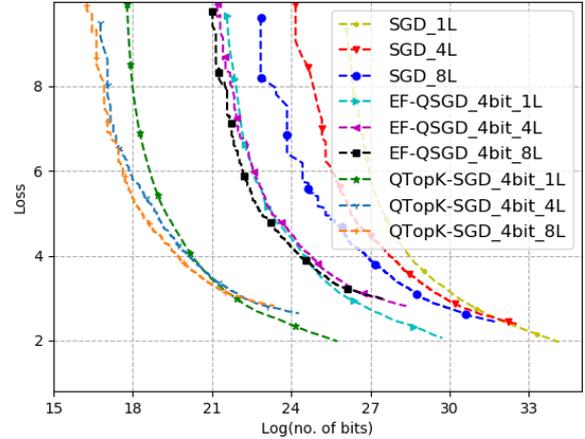
(c) Training loss vs epochs



(d) Training loss vs \log_2 of communication budget



(e) Training loss vs epochs



(f) Training loss vs \log_2 of communication budget

Figure 5 Figure 5a-5b demonstrate the effect of incorporating local iterations and compare these effects across vanilla SGD, the sparsifier $TopK$, as well as its composition with the $Sign$ operator. Similar comparisons are also made between vanilla SGD, the quantizer QSGD with error accumulation, as well as its composition with the $TopK$ sparsifier.

random from $[H]$ after every synchronization by that worker. This ensures that $gap(\mathcal{I}_T^{(r)}) \leq H$ holds for every worker $r \in [R]$ and the schedule $\mathcal{I}_T^{(r)}$ is different for each of them.

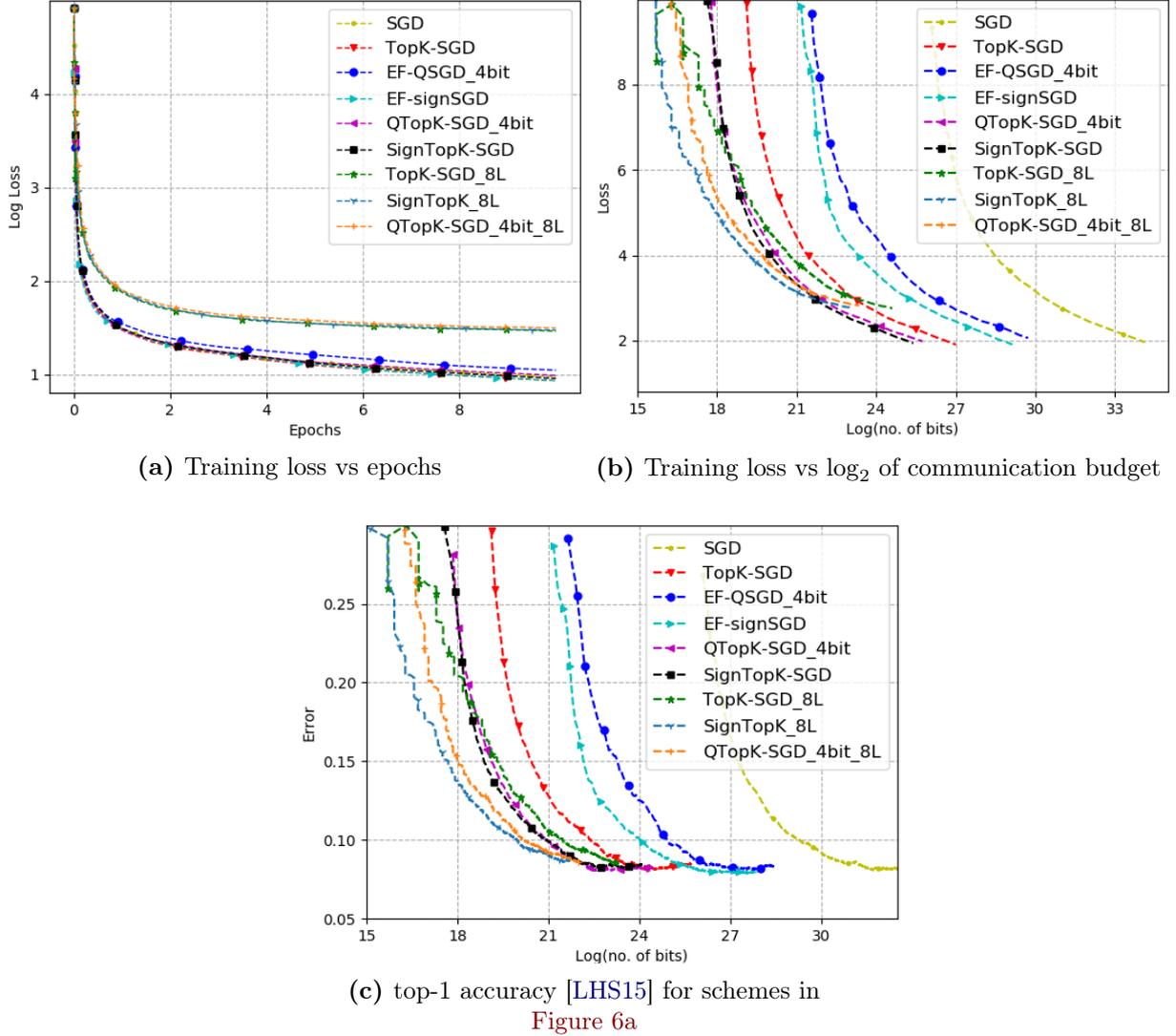
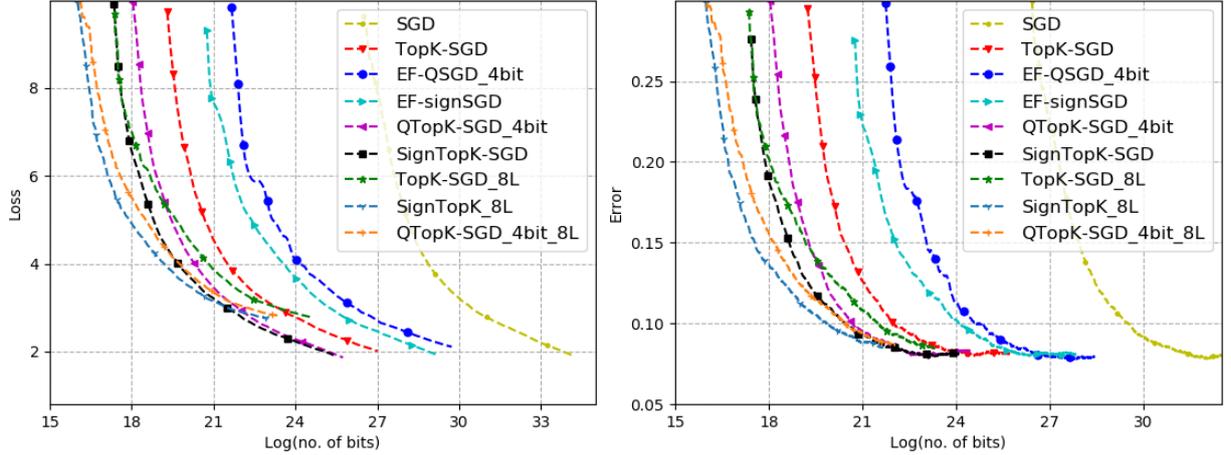


Figure 6 Figure 6a-6c demonstrate the performance of our scheme in comparison with EF-QSGD, EF-SIGNSGD [KRSJ19] and TopK-SGD [SCJ18, AHJ⁺18] in a convex setting for synchronous updates.

6 Conclusion

In this paper, we propose a gradient compression scheme that composes both unbiased and biased quantization with aggressive sparsification. Furthermore, we incorporate local computations, which, when combined with quantization and explicit sparsification, results in a highly communication efficient distributed algorithm, which we call *Qsparse-local-SGD*. We developed convergence analyses of our scheme in both synchronous as well as asynchronous settings and for both convex and non-convex objectives, and we show that our proposed algorithm achieves the same rate as that of distributed vanilla SGD in each of these cases. Our schemes provide flexibility in terms of different options for mitigating the communication bottlenecks that arise in training high-dimensional learning models over bandwidth limited networks. When run without compression, this also subsumes/generalizes several recent results from the literature on local SGD, with similar convergence rates, as mentioned at the end of Section 3.3.

Our numerics incorporate momentum acceleration, whose analysis is a topic for future research (*e.g.*, potentially by incorporating ideas from [YJY19]). Although we use momentum for



(a) Training loss with the communication budget for our schemes against baselines

(b) Test error using a model trained for given number of iterations, as seen in Figure 7a

Figure 7 Figure 7a-7b demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRSJ19] and TopK-SGD [SCJ18, AHJ⁺18] in a convex setting for asynchronous operation.

each local iteration, our preliminary results suggest that our method works with momentum applied to a block of updates as well though it was not the main focus of this paper.

Acknowledgement

The authors gratefully thank Navjot Singh for his help with experiments in the early stages of this work. This work was partially supported by NSF grant #1514531, by UC-NL grant LFR-18-548554 and by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

A Omitted Details from Section 2

A.1 Proof of Lemma 1

Lemma (Restating Lemma 1). *Let $Comp_k \in \{Top_k, Rand_k\}$. Let $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a quantizer with parameter s that satisfies Definition 1. Let $Q_s Comp_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined as $Q_s Comp_k(\mathbf{x}) := Q_s(Comp_k(\mathbf{x}))$ for every $\mathbf{x} \in \mathbb{R}^d$. If k, s are such that $\beta_{k,s} < 1$. then $Q_s Comp_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a compression operator with the compression coefficient being equal to $\gamma = (1 - \beta_{k,s}) \frac{k}{d}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s Comp_k(\mathbf{x})\|_2^2] \leq \left[1 - (1 - \beta_{k,s}) \frac{k}{d} \right] \|\mathbf{x}\|_2^2,$$

where expectation is taken over the randomness of the compression operator $Comp_k$ as well as the quantizer Q_s .

Proof. Fix an arbitrary $\mathbf{x} \in \mathbb{R}^d$.

$$\begin{aligned}
& \mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \\
&= \mathbb{E}_{C,Q}[\|\mathbf{x}\|_2^2] + \mathbb{E}_{C,Q}[\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \\
&\quad - 2\mathbb{E}_C[\langle \mathbf{x}, \mathbb{E}_Q[Q_s \text{Comp}_k(\mathbf{x})] \rangle] \\
&= \|\mathbf{x}\|_2^2 + \mathbb{E}_{C,Q}[\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] - 2\mathbb{E}_C[\langle \mathbf{x}, \text{Comp}_k(\mathbf{x}) \rangle]
\end{aligned}$$

In the last equality, we used that \mathbf{x} is constant with respect to the randomness of Q_s and Comp_k , and that $\mathbb{E}_Q[Q_s \text{Comp}_k(\mathbf{x})] = \text{Comp}_k(\mathbf{x})$, which follows from (i) of Definition 1. Observe that, for any $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$, we have $\langle \mathbf{x}, \text{Comp}_k(\mathbf{x}) \rangle = \|\text{Comp}_k(\mathbf{x})\|_2^2$. Continuing from above, we get

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s \text{Comp}_k(\mathbf{x})\|_2^2] &= \|\mathbf{x}\|_2^2 - 2\mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
&\quad + \mathbb{E}_{C,Q}[\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2]
\end{aligned} \tag{23}$$

Observe that for any $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$, $\text{Comp}_k(\mathbf{x})$ is a length- d vector, but only (at most) k of its components are non-zero. This implies that, by treating $\text{Comp}_k(\mathbf{x})$ a length- k vector whose entries correspond to the k non-zero entries of \mathbf{x} , we can write $\mathbb{E}_Q[\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \leq (1 + \beta_{k,s})\|\text{Comp}_k(\mathbf{x})\|_2^2$; see (ii) of Definition 1. Putting this back in (23), we get

$$\begin{aligned}
& \mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \\
&\leq \|\mathbf{x}\|_2^2 - \mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2] + \beta_{k,s}\mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
&= \|\mathbf{x}\|_2^2 - (1 - \beta_{k,s})\mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2]
\end{aligned} \tag{24}$$

Using $\mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2] \geq \frac{k}{d}\|\mathbf{x}\|_2^2$ (see (27) in Lemma 13) in (24) gives

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s \text{Comp}_k(\mathbf{x})\|_2^2] &\leq \|\mathbf{x}\|_2^2 - (1 - \beta_{k,s})\frac{k}{d}\|\mathbf{x}\|_2^2 \\
&= \left[1 - (1 - \beta_{k,s})\frac{k}{d}\right]\|\mathbf{x}\|_2^2.
\end{aligned}$$

This completes the proof of Lemma 1. \square

A.2 Proof of Lemma 2

Lemma (Restating Lemma 2). *Let $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a stochastic quantizer with parameter s that satisfies Definition 1. Let $Q_s \text{Comp}_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be defined as $Q_s \text{Comp}_k(\mathbf{x}) := Q_s(\text{Comp}_k(\mathbf{x}))$ for every $\mathbf{x} \in \mathbb{R}^d$. Then $\frac{Q_s \text{Comp}_k(\mathbf{x})}{1 + \beta_{k,s}}$ is a compression operator with the compression coefficient being equal to $\gamma = \frac{k}{d(1 + \beta_{k,s})}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$*

$$\mathbb{E}_{C,Q} \left[\left\| \mathbf{x} - \frac{Q_s \text{Comp}_k(\mathbf{x})}{1 + \beta_{k,s}} \right\|_2^2 \right] \leq \left[1 - \frac{k}{d(1 + \beta_{k,s})} \right] \|\mathbf{x}\|_2^2,$$

Proof. Fix an arbitrary $\mathbf{x} \in \mathbb{R}^d$.

$$\begin{aligned}
\mathbb{E}_{C,Q} \left[\left\| \mathbf{x} - \frac{Q_s \text{Comp}_k(\mathbf{x})}{(1 + \beta_{k,s})} \right\|_2^2 \right] &= \|\mathbf{x}\|_2^2 - 2\mathbb{E}_C \left[\left\langle \mathbf{x}, \mathbb{E}_Q \left[\frac{Q_s \text{Comp}_k(\mathbf{x})}{(1 + \beta_{k,s})} \right] \right\rangle \right] + \mathbb{E}_{C,Q} \left[\frac{\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2}{(1 + \beta_{k,s})^2} \right] \\
&\stackrel{(a)}{=} \|\mathbf{x}\|_2^2 - \frac{2}{(1 + \beta_{k,s})} \mathbb{E}_C [\langle \mathbf{x}, \text{Comp}_k(\mathbf{x}) \rangle]
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{(1 + \beta_{k,s})^2} \mathbb{E}_{C,Q} [\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \\
\stackrel{(b)}{=} & \|\mathbf{x}\|_2^2 - \frac{2}{(1 + \beta_{k,s})} \mathbb{E}_C [\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
& + \frac{1}{(1 + \beta_{k,s})^2} \mathbb{E}_{C,Q} [\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \\
\stackrel{(c)}{\leq} & \|\mathbf{x}\|_2^2 - \frac{2}{1 + \beta_{k,s}} \mathbb{E}_C [\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
& + \frac{1}{(1 + \beta_{k,s})} \mathbb{E}_C [\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
= & \|\mathbf{x}\|_2^2 - \frac{1}{(1 + \beta_{k,s})} \mathbb{E}_C [\|\text{Comp}_k(\mathbf{x})\|_2^2] \\
\stackrel{(d)}{\leq} & \left[1 - \frac{k}{d(1 + \beta_{k,s})} \right] \|\mathbf{x}\|_2^2. \tag{25}
\end{aligned}$$

In (a) we used $\mathbb{E}_Q[Q_s \text{Comp}_k(\mathbf{x})] = \text{Comp}_k(\mathbf{x})$, in (b) we used $\langle \mathbf{x}, \text{Comp}_k(\mathbf{x}) \rangle = \|\text{Comp}_k(\mathbf{x})\|_2^2$; in (c) we used $\mathbb{E}_Q[\|Q_s \text{Comp}_k(\mathbf{x})\|_2^2] \leq (1 + \beta_{k,s}) \|\text{Comp}_k(\mathbf{x})\|_2^2$; and in (d) we used $\mathbb{E}_C[\|\text{Comp}_k(\mathbf{x})\|_2^2] \geq \frac{k}{d} \|\mathbf{x}\|_2^2$. This completes the proof of [Lemma 2](#). \square

A.3 Proof of [Lemma 3](#)

Lemma (Restating [Lemma 3](#)). *For $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$, $\frac{\|\text{Comp}_k(\mathbf{x})\|_m \text{SignComp}_k(\mathbf{x})}{k}$, for any $m \in \mathbb{Z}_+$ is a compression operator with the compression coefficient γ_m being equal to*

$$\gamma_m = \begin{cases} \max \left\{ \frac{1}{d}, \frac{k}{d} \left(\frac{\|\text{Comp}_k(\mathbf{x})\|_1}{\sqrt{d} \|\text{Comp}_k(\mathbf{x})\|_2} \right)^2 \right\} & \text{if } m = 1, \\ \frac{k^{\frac{2}{m}-1}}{d} & \text{if } m \geq 2. \end{cases}$$

For proving [Lemma 3](#) we first state and prove [Lemma 13](#) below.

Lemma 13. *Let $\text{Comp}_k \in \{\text{Top}_k, \text{Rand}_k\}$. For any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}[\|\text{Comp}_k(\mathbf{x})\|_1^2] \geq \max \left\{ \frac{k}{d} \|\mathbf{x}\|_2^2, \frac{k^2}{d^2} \|\mathbf{x}\|_1^2 \right\} \tag{26}$$

$$\mathbb{E}[\|\text{Comp}_k(\mathbf{x})\|_2^2] \geq \frac{k}{d} \|\mathbf{x}\|_2^2. \tag{27}$$

Proof. Let $m \in \{1, 2\}$. Observe that for any $\mathbf{x} \in \mathbb{R}^d$, we have $\mathbb{E}[\|\text{Top}_k(\mathbf{x})\|_m^2] = \|\text{Top}_k(\mathbf{x})\|_m^2$ and that $\|\text{Top}_k(\mathbf{x})\|_m^2 \geq \mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_m^2]$. So, in order to prove the lemma, it suffices to show that $\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_m^2] \geq \frac{k}{d} \|\mathbf{x}\|_m^2$ holds for any $m \in \{1, 2\}$, and that $\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_1^2] \geq \frac{k^2}{d^2} \|\mathbf{x}\|_1^2$. Let Ω_k be the set of all the k -elements subsets of $[d]$.

$$\begin{aligned}
\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_m^2] &= \sum_{\omega \in \Omega_k} \frac{1}{|\Omega_k|} \left(\sum_{i=1}^d |x_i|^m \cdot \mathbb{1}\{i \in \omega\} \right)^{2/m} \\
&\stackrel{(a)}{\geq} \sum_{\omega \in \Omega_k} \frac{1}{|\Omega_k|} \sum_{i=1}^d |x_i|^2 \cdot \mathbb{1}\{i \in \omega\} \\
&= \sum_{i=1}^d x_i^2 \cdot \frac{1}{|\Omega_k|} \sum_{\omega \in \Omega_k} \mathbb{1}\{i \in \omega\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^d x_i^2 \cdot \frac{1}{|\Omega_k|} \binom{d-1}{k-1} \\
&= \frac{k}{d} \|\mathbf{x}\|_2^2
\end{aligned}$$

Note that (a) holds only for $m \in \{1, 2\}$, and it is equality for $m = 2$. Now we show that $\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_1^2] \geq \frac{k^2}{d^2} \|\mathbf{x}\|_1^2$.

$$\begin{aligned}
\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_1^2] &\geq (\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_1])^2 \\
&= \left(\sum_{\omega \in \Omega_k} \frac{1}{|\Omega_k|} \sum_{i=1}^d |x_i| \cdot \mathbb{1}\{i \in \omega\} \right)^2 \\
&= \left(\sum_{i=1}^d |x_i| \cdot \frac{1}{|\Omega_k|} \sum_{\omega \in \Omega_k} \mathbb{1}\{i \in \omega\} \right)^2 \\
&= \left(\sum_{i=1}^d |x_i| \cdot \frac{1}{|\Omega_k|} \binom{d-1}{k-1} \right)^2 \\
&= \frac{k^2}{d^2} \|\mathbf{x}\|_1^2
\end{aligned}$$

This completes the proof of [Lemma 13](#). □

Proof of Lemma 3. Fix an arbitrary $\mathbf{x} \in \mathbb{R}^d$ and consider the following:

$$\begin{aligned}
&\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \text{SignComp}_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \left\langle \frac{\|Comp_k(\mathbf{x})\|_m \text{SignComp}_k(\mathbf{x})}{k}, \mathbf{x} \right\rangle + \|\mathbf{x}\|_2^2 \right] \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \frac{\|Comp_k(\mathbf{x})\|_m \|Comp_k(\mathbf{x})\|_1}{k} + \|\mathbf{x}\|_2^2 \right] \\
&\leq \|\mathbf{x}\|_2^2 - \frac{\mathbb{E}_C \|Comp_k(\mathbf{x})\|_m^2}{k}
\end{aligned} \tag{28}$$

In (28) we used the fact that $\|\cdot\|_1 \geq \|\cdot\|_m$ for every $m \geq 1$.

Case 1. When $m = 1$: Substituting $\mathbb{E}_C \|Comp_k(\mathbf{x})\|_1^2 \geq \max \left\{ \frac{k}{d} \|\mathbf{x}\|_2^2, \frac{k^2}{d^2} \|\mathbf{x}\|_1^2 \right\}$ (from (26)) in (28) gives

$$\begin{aligned}
\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_1 \text{SignComp}_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 &\leq \|\mathbf{x}\|_2^2 - \frac{1}{k} \max \left\{ \frac{k}{d} \|\mathbf{x}\|_2^2, \frac{k^2}{d^2} \|\mathbf{x}\|_1^2 \right\} \\
&\leq \left[1 - \max \left\{ \frac{1}{d}, \frac{k}{d} \left(\frac{\|Comp_k(\mathbf{x})\|_1}{\sqrt{d} \|Comp_k(\mathbf{x})\|_2} \right)^2 \right\} \right] \|\mathbf{x}\|_2^2.
\end{aligned}$$

Case 2. When $m \geq 2$: Since $\|\mathbf{u}\|_p \leq k^{\frac{1}{p}-\frac{1}{q}} \|\mathbf{u}\|_q$ holds for every $\mathbf{u} \in \mathbb{R}^k$, whenever $p \leq q$, using

this in (28) with $q = m$ and $p = 2$ gives

$$\begin{aligned}
\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \text{Sign}Comp_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 & \\
&\leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} \mathbb{E}_C [\|Comp_k(\mathbf{x})\|_2^2] \\
&\leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} (k/d) \|\mathbf{x}\|_2^2 \quad (\text{By Lemma 13}) \\
&= \left[1 - \frac{k^{\frac{2}{m}-1}}{d} \right] \|\mathbf{x}\|_2^2. \tag{29}
\end{aligned}$$

This completes the proof of Lemma 3. \square

B Omitted Details from Section 3

B.1 Proof of Lemma 4

Lemma (Restating Lemma 4). *Let $\text{gap}(\mathcal{I}_T) \leq H$ and $\eta_t = \frac{\xi}{a+t}$, where ξ is a constant and $a > \frac{4H}{\gamma}$. Then there exists a constant $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$, such that the following holds for every worker $r \in [R]$ and for every $t \in \mathbb{Z}^+$:*

$$\mathbb{E} \|m_t^{(r)}\|_2^2 \leq 4 \frac{\eta_t^2}{\gamma^2} C H^2 G^2.$$

Proof. Fix an arbitrary worker $r \in [R]$. In order to prove the lemma, we need to show that $\mathbb{E} \|m_t^{(r)}\|_2^2 \leq 4 \frac{\eta_t^2}{\gamma^2} C H^2 G^2$ holds for every $t \in [T]$, where $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$. We show this separately for two cases, depending on whether or not $t \in \mathcal{I}_T$. First consider the case when $t \in \mathcal{I}_T$. Let $\mathcal{I}_T = \{t_{(1)}, t_{(2)}, \dots, t_{(l)} = T\}$. Fix any $i = 1, 2, \dots, l$ and consider $\mathbb{E} \|m_{t_{(i+1)}}^{(r)}\|_2^2$. Note that local memory $m_t^{(r)}$ at any worker r and the global parameter vector \mathbf{x}_t do not change in between the synchronization indices. We define $m_{t_{(0)}}^{(r)} := \mathbf{0}$ for every $r \in [R]$.

$$\begin{aligned}
\mathbb{E} \|m_{t_{(i+1)}}^{(r)}\|_2^2 &= \mathbb{E} \|m_{t_{(i+1)}-1}^{(r)} + \mathbf{x}_{t_{(i+1)}-1} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)} - g_{t_{(i+1)}-1}^{(r)}\|_2^2 \\
&\stackrel{(a)}{\leq} (1-\gamma) \mathbb{E} \|m_{t_{(i+1)}-1}^{(r)} + \mathbf{x}_{t_{(i+1)}-1} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|_2^2 \\
&\stackrel{(b)}{=} (1-\gamma) \mathbb{E} \|m_{t_{(i)}}^{(r)} + \mathbf{x}_{t_{(i)}} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|_2^2 \\
&\stackrel{(c)}{=} (1-\gamma) \mathbb{E} \|m_{t_{(i)}}^{(r)} + \widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|_2^2 \tag{30}
\end{aligned}$$

Here (a) is due to the compression property, (b) holds since the memory and master parameter remain unchanged between two rounds of synchronization, and in (c) we used that $\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} = \mathbf{x}_{t_{(i)}}$, which holds for every r . Using the inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+\tau)\|\mathbf{a}\|^2 + (1+\frac{1}{\tau})\|\mathbf{b}\|^2$, which holds for every $\tau > 0$, in (30) gives (take any $p > 1$ in the following):

$$\begin{aligned}
\mathbb{E} \|m_{t_{(i+1)}}^{(r)}\|_2^2 &\leq (1-\gamma) \left[\left(1 + \frac{(p-1)\gamma}{p} \right) \mathbb{E} \|m_{t_{(i)}}^{(r)}\|_2^2 + \left(1 + \frac{p}{(p-1)\gamma} \right) \mathbb{E} \|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|_2^2 \right] \\
&\leq \left(1 - \frac{\gamma}{p} \right) \mathbb{E} \|m_{t_{(i)}}^{(r)}\|_2^2 + \frac{(1-\gamma)(p\gamma+p)}{(p-1)\gamma} \mathbb{E} \|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|_2^2
\end{aligned}$$

$$\begin{aligned}
&= \left(1 - \frac{\gamma}{p}\right) \mathbb{E} \|m_{t(i)}^{(r)}\|^2 + \frac{p(1-\gamma^2)}{(p-1)\gamma} \mathbb{E} \left\| \widehat{\mathbf{x}}_{t(i)}^{(r)} - \widehat{\mathbf{x}}_{t(i+1)-\frac{1}{2}}^{(r)} \right\|^2 \\
&= \left(1 - \frac{\gamma}{p}\right) \mathbb{E} \|m_{t(i)}^{(r)}\|^2 + \frac{p(1-\gamma^2)}{(p-1)\gamma} \mathbb{E} \left\| \sum_{j=t(i)}^{t(i+1)-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \\
&\leq \left(1 - \frac{\gamma}{p}\right) \mathbb{E} \|m_{t(i)}^{(r)}\|^2 + \frac{p(1-\gamma^2)}{(p-1)\gamma} \eta_{t(i)}^2 H^2 G^2
\end{aligned} \tag{31}$$

In the last inequality (31) we used $\mathbb{E} \left\| \sum_{j=t(i)}^{t(i+1)-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \leq \eta_{t(i)}^2 H^2 G^2$, which can be seen as follows:

$$\begin{aligned}
\mathbb{E} \left\| \sum_{j=t(i)}^{t(i+1)-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 &= (t(i+1) - t(i))^2 \mathbb{E} \left\| \frac{1}{t(i+1) - t(i)} \sum_{j=t(i)}^{t(i+1)-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \\
&\stackrel{(a)}{\leq} (t(i+1) - t(i)) \sum_{j=t(i)}^{t(i+1)-1} \mathbb{E} \left\| \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \\
&\stackrel{(b)}{\leq} (t(i+1) - t(i)) \eta_{t(i)}^2 \sum_{j=t(i)}^{t(i+1)-1} \mathbb{E} \left\| \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \\
&\leq (t(i+1) - t(i)) \eta_{t(i)}^2 (t(i+1) - t(i)) G^2 \\
&\stackrel{(c)}{\leq} \eta_{t(i)}^2 H^2 G^2
\end{aligned}$$

Here (a) holds by Jensen's inequality, (b) holds since $\eta_t \leq \eta_{t(i)} \forall t \geq t(i)$ and (c) holds because $(t(i+1) - t(i)) \leq H$. Define $\tilde{\eta}_t = \frac{1}{a+t}$ and $A = \xi^2 H^2 G^2$. Using this in (31) gives

$$\mathbb{E} \|m_{t(i+1)}^{(r)}\|^2 \leq \left(1 - \frac{\gamma}{p}\right) \mathbb{E} \|m_{t(i)}^{(r)}\|^2 + \frac{p(1-\gamma^2)}{(p-1)\gamma} \tilde{\eta}_{t(i)}^2 A. \tag{32}$$

We want to show that $\mathbb{E} \|m_{t(i)}^{(r)}\|^2 \leq 4C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A$ holds for every $i = 1, 2, \dots$, where $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$.

In fact we prove a slightly stronger bound that $\mathbb{E} \|m_{t(i)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A$ holds for every $i = 1, 2, \dots$. We prove this using induction on i .

Base case ($i = 1$): Note that $m_{t(1)-1}^{(r)} = m_0^{(r)} = \mathbf{0}$. Consider the following:

$$\begin{aligned}
\mathbb{E} \|m_{t(1)}^{(r)}\|^2 &= \mathbb{E} \left\| \mathbf{x}_{t(1)-1} - \widehat{\mathbf{x}}_{t(1)-\frac{1}{2}} - g_{t(1)-1}^{(r)} \right\|^2 \\
&\leq (1 - \gamma) \mathbb{E} \left\| \mathbf{x}_{t(1)-1} - \widehat{\mathbf{x}}_{t(1)-\frac{1}{2}} \right\|^2 \\
&\stackrel{(a)}{=} (1 - \gamma) \mathbb{E} \left\| \widehat{\mathbf{x}}_0^{(r)} - \widehat{\mathbf{x}}_{t(1)-\frac{1}{2}} \right\|^2 \\
&= (1 - \gamma) \mathbb{E} \left\| \sum_{j=0}^{t(1)-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \right\|^2 \\
&\leq (1 - \gamma) \eta_0^2 H^2 G^2 \\
&= (1 - \gamma) \tilde{\eta}_0^2 A
\end{aligned}$$

Here (a) holds since $\mathbf{x}_{t(1)-1} = \mathbf{x}_0 = \widehat{\mathbf{x}}_0^{(r)}$. It is easy to verify that $(1 - \gamma) \tilde{\eta}_0^2 A \leq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H} \frac{\tilde{\eta}_{t(1)}^2}{\gamma^2} A$. To show this, we use $\frac{\tilde{\eta}_0}{\tilde{\eta}_{t(1)}} = \frac{a+t(1)}{a} \leq \frac{a+H}{a} \leq 2$, where the first inequality follows from $t(1) \leq H$

and the second inequality follows from $a \geq H$. Now, since $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$, it follows that $\mathbb{E}\|m_{t(1)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t(1)}^2}{\gamma^2} A$.

Inductive case: Assume $\mathbb{E}\|m_{t(i)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A$ for some $i \in \mathbb{Z}^+$. We need to show that $\mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t(i+1)}^2}{\gamma^2} A$. Using the inductive hypothesis in (32), we get

$$\begin{aligned} \mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 &\leq \left(1 - \frac{\gamma}{p}\right) C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A + \frac{p(1-\gamma^2)}{(p-1)\gamma} \tilde{\eta}_{t(i)}^2 A \\ &= C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A \left(1 - \frac{\gamma}{p} + \frac{p(1-\gamma^2)}{p-1} \frac{\gamma}{C}\right) \\ &= C \frac{\tilde{\eta}_{t(i)}^2}{\gamma^2} A \left(1 - \frac{\gamma}{p} \left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right)\right) \end{aligned} \quad (33)$$

Claim 1. For any $p > 1$, if $\frac{\gamma}{p} \left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right) \geq \frac{2H}{a}$, then $\tilde{\eta}_{t(i)}^2 \left(1 - \frac{\gamma}{p} \left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right)\right) \leq \tilde{\eta}_{t(i+1)}^2$ holds.

Proof. Let $\frac{\gamma}{p} \left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right) = \frac{\beta}{a}$. Since $t_{(i+1)} \leq t_{(i)} + H$ (which implies that $\tilde{\eta}_{t_{(i)}+H}^2 \leq \tilde{\eta}_{t_{(i+1)}}^2$), it suffices to show that $\tilde{\eta}_{t_{(i)}}^2 \left(1 - \frac{\beta}{a}\right) \leq \tilde{\eta}_{t_{(i)}+H}^2$ holds whenever $\beta \geq 2H$. For simplicity of notation, let $t = t_{(i)}$. Note that $\tilde{\eta}_t^2 \left(1 - \frac{\beta}{a}\right) = \frac{(a-\beta)}{a(a+t)^2}$. We show below that if $\beta > 2H$, then $a(a+t)^2 \geq (a+t+H)^2(a-\beta)$. This proves our claim, because now we have $\frac{(a-\beta)}{a(a+t)^2} \leq \frac{(a-\beta)}{(a+t+H)^2(a-\beta)} = \frac{1}{(a+t+H)^2} = \tilde{\eta}_{t+H}^2$. It only remains to show that $a(a+t)^2 \leq (a+t+H)^2(a-\beta)$ holds if $\beta \geq 2H$.

$$\begin{aligned} (a+t+H)^2(a-\beta) &= ((a+t)^2 + H^2 + 2H(a+t))(a-\beta) \\ &= a(a+t)^2 + aH^2 + 2Ha^2 + 2Hat - \beta(a+t)^2 - \beta H^2 - 2H\beta(a+t) \\ &= a(a+t)^2 + a(H^2 + 2Ht - 2\beta t - 2H\beta) + a^2(2H - \beta) \\ &\quad - \beta t^2 - \beta H^2 - 2H\beta t \\ &\leq a(a+t)^2. \end{aligned}$$

The last inequality holds whenever $\beta \geq 2H$. \square

Therefore we need $\frac{\gamma}{p} \left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right) \geq \frac{2H}{a}$, which is equivalent to requiring $C \geq \frac{\gamma a p^2(1-\gamma^2)}{(p-1)(a\gamma-2pH)}$, where $a > \frac{2pH}{\gamma}$. Since this holds for every $p > 1$, by substituting $p = 2$, we get $C \geq \frac{4\gamma a(1-\gamma^2)}{(a\gamma-4H)}$. This together with (33) and Claim 1 implies that if $C \geq \frac{4\gamma a(1-\gamma^2)}{(a\gamma-4H)}$, where $a > 4H/\gamma$, then $\mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t(i+1)}^2}{\gamma^2} A$ holds. This proves our inductive step.

We have shown that $\mathbb{E}\|m_t^{(r)}\|^2 \leq 4C \frac{\tilde{\eta}_t^2}{\gamma^2} A$ holds when $t \in \mathcal{I}_T$. It only remains to show that $\mathbb{E}\|m_t^{(r)}\|^2 \leq 4C \frac{\tilde{\eta}_t^2}{\gamma^2} A$ also holds when $t \in [T] \setminus \mathcal{I}_T$. Let $i \in \mathbb{Z}_+$ be such that $t_{(i)} \leq t < t_{(i+1)}$, which implies that $\tilde{\eta}_{t_{(i)}} \leq 2\tilde{\eta}_t$. Since local memory does not change in between the synchronization indices, we have that $m_t^{(r)} = m_{t_{(i)}}^{(r)}$. Thus we have $\mathbb{E}\|m_t^{(r)}\|^2 = \mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t_{(i)}}^2}{\gamma^2} A \leq 4C \frac{\tilde{\eta}_t^2}{\gamma^2} A$. This concludes the proof of Lemma 4. \square

B.2 Proof of Lemma 5

Lemma (Restating Lemma 5). *Let $\text{gap}(\mathcal{I}_T) \leq H$. Then the following holds for every worker $r \in [R]$ and for every $t \in \mathbb{Z}^+$:*

$$\mathbb{E}\|m_t^{(r)}\|_2^2 \leq 4 \frac{\eta^2(1-\gamma^2)}{\gamma^2} H^2 G^2.$$

Proof. Observe that (31) holds irrespective of the learning rate schedule. In particular, using a fixed learning rate $\eta_t = \eta$ for every t gives

$$\mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 \leq \left(1 - \frac{\gamma}{p}\right) \mathbb{E}\|m_{t(i)}^{(r)}\|^2 + \frac{p(1-\gamma^2)}{(p-1)\gamma} \eta^2 H^2 G^2$$

When rolled out we see that the memory is upper bounded by a geometric sum.

$$\begin{aligned} \mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 &\leq \frac{p(1-\gamma^2)}{(p-1)\gamma} \eta^2 H^2 G^2 \sum_{j=0}^{\infty} \left(1 - \frac{\gamma}{p}\right)^j \\ &\leq \frac{p^2(1-\gamma^2)}{(p-1)} \frac{\eta^2}{\gamma^2} H^2 G^2. \end{aligned}$$

Note that the last inequality holds for every $p > 1$, and is minimized when $p = 2$. By plugging $p = 2$, we get

$$\mathbb{E}\|m_{t(i+1)}^{(r)}\|^2 \leq \frac{4(1-\gamma^2)\eta^2}{\gamma^2} H^2 G^2.$$

Since the RHS does not depend on t , it follows that $\mathbb{E}\|m_t^{(r)}\|^2 \leq \frac{4(1-\gamma^2)\eta^2}{\gamma^2} H^2 G^2$ holds for every $t \in [T]$. This completes the proof of Lemma 5. \square

B.3 Proof of Lemma 6

Lemma (Restating Lemma 6). *Let $\tilde{\mathbf{x}}_t^{(r)}, m_t^{(r)}$, $r \in [R]$, $t \geq 0$ be generated according to Algorithm 1 and let $\hat{\mathbf{x}}_t^{(r)}$ be as defined in (4). Let $\tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_t^{(r)}$ and $\hat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)}$. Then we have*

$$\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)},$$

i.e., the difference of the true and the virtual sequence is equal to the average memory.

Proof. Now consider $\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)} - \tilde{\mathbf{x}}_t^{(r)}$. For the nearest $t_r + 1 \in \mathcal{I}_T$ such that $t_r + 1 \leq t$ and the nearest $t'_r + 1 \in \mathcal{I}_T$ such that $t'_r + 1 \leq t_r$

$$\begin{aligned} \hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t &= \frac{1}{R} \sum_{r=1}^R \left(\hat{\mathbf{x}}_{t_r+1}^{(r)} - \tilde{\mathbf{x}}_{t_r+1}^{(r)} \right) \\ &= \frac{1}{R} \sum_{r=1}^R \left(\mathbf{x}_{t_r} - \frac{1}{R} \sum_{r=1}^R g_{t_r}^{(r)} - (\tilde{\mathbf{x}}_{t'_r+1}^{(r)} - (\hat{\mathbf{x}}_{t'_r+1}^{(r)} - \hat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)})) \right) \end{aligned} \quad (34)$$

Here we used that $\widehat{\mathbf{x}}_{t'_r+1}^{(r)} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)} = \sum_{j=t'_r+1}^{t_r} \eta_j \nabla^{(r)} f_{(i_j)} \left(\widehat{\mathbf{x}}_j^{(r)} \right)$. Substituting $\widehat{\mathbf{x}}_{t'_r+1}^{(r)} = \mathbf{x}_{t'_r+1}$ we get

$$\begin{aligned} \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R} \sum_{r=1}^R \left(\mathbf{x}_{t_r} - \frac{1}{R} \sum_{r=1}^R g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t'_r+1}^{(r)} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)})) \right) \\ &= \mathbf{x}_{t'_r+1} - \frac{1}{R} \sum_{r=1}^R g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t'_r+1}^{(r)} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)})) \\ &= \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)}) - \frac{1}{R} \sum_{r=1}^R g_{t_r}^{(r)} \end{aligned} \quad (35)$$

Now since $\mathbf{x}_{t'_r+1} = \mathbf{x}_{t_r}$ we have

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t_r} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)}) - \frac{1}{R} \sum_{r=1}^R g_{t_r}^{(r)} \quad (36)$$

On rolling out the expression in (36) we get

$$\begin{aligned} \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R} \sum_{r=1}^R \left[\sum_{\substack{j:j+1 \in \mathcal{I}_T \\ j \leq t_r}} \left(\mathbf{x}_j^{(r)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(r)} - g_j^{(r)} \right) \right] \\ &= \frac{1}{R} \sum_{r=1}^R m_{t_r+1}^{(r)} \\ &= \frac{1}{R} \sum_{r=1}^R m_t^{(r)} \end{aligned} \quad (37)$$

Therefore $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)}$ is the average memory. This completes the proof of [Lemma 6](#). \square

B.4 Proof of [Lemma 7](#)

Lemma (Restating [Lemma 7](#)). *Let $\text{gap}(\mathcal{I}_T) \leq H$. For $\widehat{\mathbf{x}}_t^{(r)}$ generated according to [Algorithm 1](#) with a fixed learning rate η and letting $\widehat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$, we have the following bound on the deviation of the local sequences:*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq \eta^2 G^2 H^2.$$

Proof. To prove this, we follow the proof of [Lemma 8](#) until (38) and put $\eta_{t_r} = \eta$ to get $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq \eta^2 G^2 H^2$. \square

B.5 Proof of [Lemma 8](#)

Lemma (Restating [Lemma 8](#)). *Let $\text{gap}(\mathcal{I}_T) \leq H$. By running [Algorithm 1](#) with a decaying learning rate η_t , we have*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq 4\eta_t^2 G^2 H^2.$$

Proof. We show this along the lines of the proof of [Sti19, Lemma 3.3]. We need to upper-bound $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2$. Note that for any R vectors $\mathbf{u}_1, \dots, \mathbf{u}_R$, if we let $\bar{\mathbf{u}} = \frac{1}{R} \sum_{i=1}^R \mathbf{u}_i$, then $\sum_{i=1}^R \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 \leq \sum_{i=1}^R \|\mathbf{u}_i\|^2$. We use this in the first inequality below.

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &= \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)} - (\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_{t_r}^{(r)})\|^2 \\ &\leq \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \\ &\leq \eta_{t_r}^2 G^2 H^2 \tag{38} \\ &\leq 4\eta_t^2 G^2 H^2 \tag{39} \end{aligned}$$

The last inequality (39) uses $\eta_{t_r} \leq 2\eta_{t_r+H} \leq 2\eta_t$ and $t - t_r \leq H$. \square

B.6 Proof of Theorem 1

Proof. Let \mathbf{x}^* be the minimizer of $f(\mathbf{x})$, therefore we denote $f(\mathbf{x}^*)$ by f^* . For the purpose of reusing the proof later while proving Theorem 2, we start off with the decaying learning rate η_t until (43) and then switch to the fixed learning rate η . Note that the proof remains the same until (43) irrespective of the learning rate schedule; in particular, we can take $\eta_t = \eta$ and the same proof holds until (43).

By the definition of L -smoothness, we have

$$\begin{aligned} f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) &\leq \langle \nabla f(\tilde{\mathbf{x}}_t), \tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t \rangle + \frac{L}{2} \|\tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t\|^2 \\ &= -\eta_t \langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{p}_t \rangle + \frac{\eta_t^2 L}{2} \|\mathbf{p}_t\|^2 \\ &= -\eta_t \langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{p}_t \rangle + \frac{\eta_t^2 L}{2} \|\mathbf{p}_t - \bar{\mathbf{p}}_t + \bar{\mathbf{p}}_t\|^2 \\ &\leq -\eta_t \langle \nabla f(\tilde{\mathbf{x}}_t), \mathbf{p}_t \rangle + \eta_t^2 L \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|^2 + \eta_t^2 L \|\bar{\mathbf{p}}_t\|^2 \quad (\text{Using Jensen's Inequality}) \\ &= -\frac{\eta_t}{R} \sum_{r=1}^R \langle \nabla f(\tilde{\mathbf{x}}_t), \nabla f_{i_t^{(r)}}(\widehat{\mathbf{x}}_t^{(r)}) \rangle + \eta_t^2 L \left\| \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)}) \right\|^2 + \eta_t^2 L \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|^2 \end{aligned}$$

Define i_t as the set of random sampling of the mini-batches at each worker $\{i_t^{(1)}, i_t^{(2)}, \dots, i_t^{(R)}\}$. Taking expectation w.r.t. the sampling at time t (conditioned on the past) and using the Lipschitz continuity of the gradients of local functions gives

$$\begin{aligned} \mathbb{E}_{i_t} [f(\tilde{\mathbf{x}}_{t+1})] - f(\tilde{\mathbf{x}}_t) &\leq -\frac{\eta_t}{2} \left(\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + \left\| \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)}) \right\|^2 - \left\| \nabla f(\tilde{\mathbf{x}}_t) - \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)}) \right\|^2 \right) \\ &\quad + \eta_t^2 L \left\| \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)}) \right\|^2 + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 \\ &\leq -\frac{\eta_t}{2R} \sum_{r=1}^R \left(\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 - L^2 \|\tilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \right) + \frac{2\eta_t^2 L - \eta_t}{2} \left\| \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)}) \right\|^2 \\ &\quad + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 \end{aligned}$$

$$\begin{aligned}
&= -\frac{\eta_t}{2R} \sum_{r=1}^R \left(\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 + L^2 \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \right) + \frac{2\eta_t^2 L - \eta_t}{2R} \sum_{r=1}^R \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 \\
&\quad + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + \frac{\eta_t L^2}{R} \sum_{r=1}^R \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2. \tag{40}
\end{aligned}$$

We bound the first term in terms of $\|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2$ as follows:

$$\begin{aligned}
\|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 &\leq 2\|\nabla f(\hat{\mathbf{x}}_t^{(r)}) - \nabla f(\tilde{\mathbf{x}}_t)\|^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|^2 \\
&\leq 2L^2 \|\hat{\mathbf{x}}_t^{(r)} - \tilde{\mathbf{x}}_t\|^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|^2, \tag{41}
\end{aligned}$$

where the 2nd inequality follows from the smoothness (L -Lipschitz gradient) assumption. Using this and that $\eta_t \leq \frac{1}{2L}$ in (40) and rearranging terms give

$$\frac{\eta_t}{4R} \sum_{r=1}^R \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 \leq f(\tilde{\mathbf{x}}_t) - \mathbb{E}_{(i_t)}[f(\tilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + \frac{\eta_t L^2}{R} \sum_{r=1}^R \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{42}$$

Taking expectation w.r.t. to the entire process and using the inequality $\|\mathbf{u} + \mathbf{v}\|^2 \leq 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$ gives

$$\begin{aligned}
\frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 2\eta_t L^2 \mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|^2 \\
&\quad + 2\eta_t L^2 \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{43}
\end{aligned}$$

Observe that (43) holds irrespective of the learning rate schedule. In particular, if we take a fixed learning rate $\eta_t = \eta \leq \frac{1}{2L}$ in (43), we get

$$\begin{aligned}
\frac{\eta}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})] + \frac{\eta^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 2\eta L^2 \mathbb{E} \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|^2 \\
&\quad + 2\eta L^2 \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{44}
\end{aligned}$$

Lemma 6 and **Lemma 5** together imply $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 \leq \frac{4\eta^2(1-\gamma^2)}{\gamma^2} G^2 H^2$. We also have from **Lemma 7** that $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \leq \eta^2 G^2 H^2$. Substituting these in (44) gives

$$\begin{aligned}
\frac{\eta}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \mathbb{E}[f(\tilde{\mathbf{x}}_{t+1})] + \frac{\eta^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 8\frac{\eta^3(1-\gamma^2)}{\gamma^2} L^2 G^2 H^2 \\
&\quad + 2\eta^3 L^2 G^2 H^2 \tag{45}
\end{aligned}$$

By taking a telescopic sum from $t = 0$ to $t = T - 1$, we get

$$\begin{aligned}
\frac{1}{4RT} \sum_{t=0}^{T-1} \sum_{r=1}^R \mathbb{E} \|\nabla f(\hat{\mathbf{x}}_t^{(r)})\|^2 &\leq \frac{\mathbb{E}[f(\tilde{\mathbf{x}}_0)] - f^*}{\eta T} + \frac{\eta L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 8\frac{\eta^2(1-\gamma^2)}{\gamma^2} L^2 G^2 H^2 \\
&\quad + 2\eta^2 L^2 G^2 H^2 \tag{46}
\end{aligned}$$

Take $\eta = \frac{\widehat{C}}{\sqrt{T}}$, where \widehat{C} is a constant (that satisfies $\widehat{C} < \frac{\sqrt{T}}{2L}$). For example, we can take $\widehat{C} = \frac{1}{2L}$. This gives

$$\frac{1}{RT} \sum_{t=0}^{T-1} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{\widehat{C}} + \frac{\widehat{C}L}{bR^2} \sum_{r=1}^R \sigma_r^2 \right) \frac{4}{\sqrt{T}} + 8 \left(4 \frac{(1-\gamma^2)}{\gamma^2} + 1 \right) \frac{\widehat{C}^2 L^2 G^2 H^2}{T}. \quad (47)$$

Sample a parameter \mathbf{z}_T from $\{\widehat{\mathbf{x}}_t^{(r)}\}$ for $r = 1, \dots, R$ and $t = 0, 1, \dots, T-1$ with probability $\Pr[\mathbf{z}_T = \widehat{\mathbf{x}}_t^{(r)}] = \frac{1}{RT}$, which implies $\mathbb{E} \|\mathbf{z}_T\|^2 = \frac{1}{RT} \sum_{t=0}^{T-1} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2$. Using this in (47) gives

$$\mathbb{E} \|\mathbf{z}_T\|^2 = \left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{\widehat{C}} + \frac{\widehat{C}L}{bR^2} \sum_{r=1}^R \sigma_r^2 \right) \frac{4}{\sqrt{T}} + 8 \left(4 \frac{(1-\gamma^2)}{\gamma^2} + 1 \right) \frac{\widehat{C}^2 L^2 G^2 H^2}{T}.$$

This completes the proof of [Theorem 1](#). \square

B.7 Proof of [Theorem 2](#)

Proof. Observe that we can use the proof of [Theorem 1](#) exactly until (43), for $\eta_t \leq \frac{1}{2L}$ (which follows from our assumption that $a \geq 2\xi L$), which gives

$$\begin{aligned} \frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + 2\eta_t L^2 \mathbb{E} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 \\ &\quad + 2\eta_t L^2 \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \end{aligned} \quad (48)$$

We have from [Lemma 8](#) that $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2$. [Lemma 6](#) and [Lemma 4](#) together imply that $\mathbb{E} \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \frac{1}{R} \sum_{r=1}^R \|m_t^{(r)}\|^2 \leq C \frac{4\eta_t^2}{\gamma^2} G^2 H^2$. Using these bounds in (48) gives

$$\frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2} \sum_{r=1}^R \sigma_r^2 + \frac{8\eta_t^3}{\gamma^2} C L^2 G^2 H^2 + 8\eta_t^3 L^2 G^2 H^2$$

Taking a telescopic sum from $t = 0$ to $t = T-1$ gives

$$\sum_{t=0}^{T-1} \frac{\eta_t}{4R} \sum_{r=1}^R \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\mathbf{x}_0)] - f^* + \frac{L \sum_{r=1}^R \sigma_r^2}{bR^2} \sum_{t=0}^{T-1} \eta_t^2 + \left(\frac{8C}{\gamma^2} + 8 \right) L^2 G^2 H^2 \sum_{t=0}^{T-1} \eta_t^3. \quad (49)$$

Let $\delta_t := \frac{\eta_t}{4R}$ and $P_T := \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t$. We show at the end of this proof that $P_T \geq \frac{\xi}{4} \ln \left(\frac{T+a-1}{a} \right)$, $\sum_{t=0}^{T-1} \eta_t^2 \leq \frac{\xi^2}{a-1}$, and that $\sum_{t=0}^{T-1} \eta_t^3 \leq \frac{\xi^3}{2(a-1)^2}$. Using these in (49) yields

$$\begin{aligned} \frac{1}{P_T} \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\leq \frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{P_T} + \frac{L\xi^2}{bR^2(a-1)} \frac{\sum_{r=1}^R \sigma_r^2}{P_T} \\ &\quad + \left(\frac{8C}{\gamma^2} + 8 \right) L^2 G^2 H^2 \frac{\xi^3}{2P_T(a-1)^2} \end{aligned} \quad (50)$$

We therefore can show a weak convergence result, i.e.,

$$\min_{t \in \{0, \dots, T-1\}, r \in [R]} \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \xrightarrow{T \rightarrow \infty} 0. \quad (51)$$

Sample a parameter \mathbf{z}_T from $\{\widehat{\mathbf{x}}_t^{(r)}\}$ for $r = 1, \dots, R$ and $t = 0, 1, \dots, T-1$ with probability $\Pr[\mathbf{z}_T = \widehat{\mathbf{x}}_t^{(r)}] = \frac{\delta_t}{P_T}$. This gives $\mathbb{E} \|\nabla f(\mathbf{z}_T)\|^2 = \frac{1}{P_T} \sum_{t=0}^{T-1} \sum_{r=1}^R \delta_t \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2$. We therefore have the following from (50)

$$\mathbb{E} \|\nabla f(\mathbf{z}_T)\|^2 \leq \frac{\mathbb{E} f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2 \sum_{r=1}^R \sigma^2}{bR^2(a-1)P_T} + \left(\frac{8C}{\gamma^2} + 8\right) \frac{\xi^3 L^2 G^2 H^2}{2(a-1)^2 P_T}$$

Since $\min_{t \in \{0, \dots, T-1\}, r \in [R]} \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2$, we have a weak convergence result:

$$\min_{t \in \{0, \dots, T-1\}, r \in [R]} \mathbb{E} \|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \xrightarrow{T \rightarrow \infty} 0.$$

Bounding the terms P_T , $\sum_{t=0}^{T-1} \eta_t^2$ and $\sum_{t=0}^{T-1} \eta_t^3$:

$$\begin{aligned} P_T &= \frac{1}{4} \sum_{t=0}^{T-1} \eta_t \geq \frac{1}{4} \sum_{t=0}^{T-1} \eta_t \geq \frac{\xi}{4} \ln \left(\frac{T+a-1}{a} \right) \\ \sum_{t=0}^{T-1} \eta_t^2 &\leq \xi^2 \left(\frac{1}{a-1} - \frac{1}{T+a-1} \right) = \frac{\xi^2 T}{(a-1)(T+a-1)} \leq \frac{\xi^2}{a-1} \\ \sum_{t=0}^{T-1} \eta_t^3 &\leq \frac{\xi^3}{2} \left(\frac{1}{(a-1)^2} - \frac{1}{(T+a-1)^2} \right) \leq \frac{\xi^3}{2(a-1)^2} \end{aligned}$$

This completes the proof of [Theorem 2](#). \square

B.8 Proof of [Theorem 3](#)

Proof. Let \mathbf{x}^* be the minimizer of $f(\mathbf{x})$, therefore we have $\nabla f(\mathbf{x}^*) = 0$. We denote $f(\mathbf{x}^*)$ by f^* . By taking the average of the virtual sequences $\widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}(\widehat{\mathbf{x}}_t^{(r)})$ for each worker $r \in [R]$ and defining $\mathbf{p}_t := \frac{1}{R} \sum_{r=1}^R \nabla f_{i_t^{(r)}}(\widehat{\mathbf{x}}_t^{(r)})$, we get

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{\mathbf{x}}_t - \eta_t \mathbf{p}_t. \quad (52)$$

Define i_t as the set of random sampling of the mini-batches at each worker $\{i_t^{(1)}, i_t^{(2)}, \dots, i_t^{(R)}\}$ and let $\bar{\mathbf{p}}_t = \mathbb{E}_{i_t}[\mathbf{p}_t]$. From (52) we can get

$$\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 + \eta_t^2 \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|^2 - 2\eta_t \langle \widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t, \mathbf{p}_t - \bar{\mathbf{p}}_t \rangle \quad (53)$$

Taking the expectation w.r.t. the sampling i_t at time t (conditioning on the past) and noting that last term in (53) becomes zero gives:

$$\mathbb{E}_{i_t} \|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 + \eta_t^2 \mathbb{E}_{i_t} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|^2 \quad (54)$$

It follows from the Jensen's inequality and independence that $\mathbb{E}_{i_t} \|\mathbf{p}_t - \bar{\mathbf{p}}_t\|^2 \leq \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$. This gives

$$\mathbb{E}_{i_t} \|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}. \quad (55)$$

Now we bound the first term on the RHS.

Lemma 14. *If $\eta_t \leq \frac{1}{4L}$, then we have*

$$\begin{aligned} \|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t\mu}{2L} (f(\tilde{\mathbf{x}}_t) - f^*) \\ &\quad + \eta_t \left(\frac{3\mu}{2} + 3L\right) \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \end{aligned} \quad (56)$$

Proof.

$$\|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 = \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 + \eta_t^2 \|\bar{\mathbf{p}}_t\|^2 - 2\eta_t \langle \tilde{\mathbf{x}}_t - \mathbf{x}^*, \bar{\mathbf{p}}_t \rangle \quad (57)$$

Using the definition of $\bar{\mathbf{p}}_t$ we have

$$\begin{aligned} \|\bar{\mathbf{p}}_t\|^2 &= \left\| \frac{1}{R} \sum_{r=1}^R \left(\nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) - \nabla f^{(r)}(\tilde{\mathbf{x}}_t) \right) + \nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*) \right\|^2 \\ &\leq \frac{1}{R} \sum_{r=1}^R 2 \|\nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) - \nabla f^{(r)}(\tilde{\mathbf{x}}_t)\|^2 + 2 \|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*)\|^2 \\ &\leq \frac{2L^2}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t^{(r)} - \tilde{\mathbf{x}}_t\| + 2 \|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*)\|^2 \end{aligned} \quad (58)$$

By the definition of smoothness, we have $\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*)\|^2 \leq 2L(f(\tilde{\mathbf{x}}_t) - f(\mathbf{x}^*))$, where $\nabla f(\mathbf{x}^*) = 0$. Substituting this in (58) gives

$$\eta_t^2 \|\bar{\mathbf{p}}_t\|^2 \leq \frac{2\eta_t^2 L^2}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t^{(r)} - \tilde{\mathbf{x}}_t\| + 4\eta_t^2 L (f(\tilde{\mathbf{x}}_t) - f(\mathbf{x}^*)) \quad (59)$$

Now we bound the last term of (57). By definition, we have

$$-2\eta_t \langle \tilde{\mathbf{x}}_t - \mathbf{x}^*, \bar{\mathbf{p}}_t \rangle = -2\frac{\eta_t}{R} \sum_{r=1}^R \langle \hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) \rangle - 2\frac{\eta_t}{R} \sum_{r=1}^R \langle \tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) \rangle \quad (60)$$

For the first term on the RHS of (60), we can use strong convexity

$$-2 \langle \hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) \rangle \leq -2 \left(f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) - f^{(r)}(\mathbf{x}^*) \right) - \mu \|\hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \quad (61)$$

For the second term on the RHS of (60), we can use the following by smoothness.

$$-2 \langle \tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) \rangle \leq L \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 + 2 \left(f^{(r)}(\hat{\mathbf{x}}_t^{(r)}) - f^{(r)}(\tilde{\mathbf{x}}_t) \right) \quad (62)$$

Using (61)-(62) in (60) we get

$$\begin{aligned} -2\eta_t \langle \tilde{\mathbf{x}}_t - \mathbf{x}^*, \bar{\mathbf{p}}_t \rangle &\leq -\frac{2\eta_t}{R} \sum_{r=1}^R \left(f^{(r)}(\tilde{\mathbf{x}}_t) - f^{(r)}(\mathbf{x}^*) \right) - \frac{\eta_t\mu}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 + \frac{L\eta_t}{R} \sum_{r=1}^R \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \\ &= -2\eta_t (f(\tilde{\mathbf{x}}_t) - f(\mathbf{x}^*)) - \frac{\eta_t\mu}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 + L \frac{\eta_t}{R} \sum_{r=1}^R \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \end{aligned} \quad (63)$$

Adding (59) and (63) and using $a \geq 32L/\mu$ which implies $\eta_t \leq 1/4L$ yields

$$\begin{aligned}
\eta_t^2 \|\bar{\mathbf{p}}_t\|^2 - 2\eta_t \langle \tilde{\mathbf{x}}_t - \mathbf{x}^*, \bar{\mathbf{p}}_t \rangle &\leq -2\eta_t(1 - 2\eta_t L) (f(\tilde{\mathbf{x}}_t) - f^*) - \frac{\eta_t \mu}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \\
&\quad + \frac{L\eta_t + 2\eta_t^2 L^2}{R} \sum_{r=1}^R \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \\
&\leq -\eta_t (f(\tilde{\mathbf{x}}_t) - f^*) - \eta_t \mu \|\hat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \frac{3L\eta_t}{R} \sum_{r=1}^R \left(\|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|^2 + \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \right) \tag{64}
\end{aligned}$$

Since $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$, we have

$$-\|\hat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \leq \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 - \frac{1}{2} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \tag{65}$$

Using (65) in (64) and then substituting (64) in (57) gives

$$\begin{aligned}
\|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \eta_t (f(\tilde{\mathbf{x}}_t) - f^*) \\
&\quad + \eta_t (\mu + 3L) \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{66}
\end{aligned}$$

Using strong convexity of f we have

$$\begin{aligned}
\|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{2} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \eta_t (\mu + 3L) \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{67}
\end{aligned}$$

Now use $-\|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \leq \|\tilde{\mathbf{x}}_t - \hat{\mathbf{x}}_t\|^2 - \frac{1}{2} \|\hat{\mathbf{x}}_t - \mathbf{x}^*\|^2$ We get

$$\begin{aligned}
\|\tilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \bar{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{4} \|\hat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \eta_t \left(\frac{3\mu}{2} + 3L\right) \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \\
&\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{2L} (f(\hat{\mathbf{x}}_t) - f^*) \quad (\text{Using smoothness of } f(\mathbf{x})) \\
&\quad + \eta_t \left(\frac{3\mu}{2} + 3L\right) \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^R \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \tag{68}
\end{aligned}$$

This completes the proof of Lemma 14. \square

Using (68) in (55) and then taking the expectation over the entire process gives

$$\begin{aligned}
\mathbb{E} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{2L} (\mathbb{E}[f(\hat{\mathbf{x}}_t)] - f^*) \\
&\quad + \eta_t \left(\frac{3\mu}{2} + 3L\right) \mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \tag{69}
\end{aligned}$$

From Lemma 8, we have $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2$. Lemma 6 and Lemma 4 together imply that $\mathbb{E} \|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 \leq 4C \frac{\eta_t^2}{\gamma^2} H^2 G^2$. Substituting these back in (69) and letting $e_t = \mathbb{E}[f(\hat{\mathbf{x}}_t) - f^*]$ gives

$$\begin{aligned} \mathbb{E} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\mu\eta_t}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) C \frac{4\eta_t^2}{\gamma^2} G^2 H^2 \\ &\quad + (3L\eta_t) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \end{aligned} \quad (70)$$

Now using $\eta_t \leq 1/4L$ we have

$$\begin{aligned} \mathbb{E} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\mu\eta_t}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) C \frac{4\eta_t^2}{\gamma^2} G^2 H^2 \\ &\quad + (3\eta_t L) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \end{aligned} \quad (71)$$

Employing a slightly modified Lemma 3.3 from [SCJ18] with $a_t = \mathbb{E} \|\tilde{\mathbf{x}}_t - \mathbf{x}^*\|^2$, $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$ and $B = 4 \left(\left(\frac{3\mu}{2} + 3L\right) \frac{CG^2H^2}{\gamma^2} + 3L^2G^2H^2 \right)$, we have

$$a_{t+1} \leq \left(1 - \frac{\mu\eta_t}{2}\right) a_t - \frac{\mu\eta_t}{2L} e_t + \eta_t^2 A + \eta_t^3 B \quad (72)$$

For $\eta_t = \frac{8}{\mu(a+t)}$ and $w_t = (a+t)^2$, $S_T = \sum_{t=0}^{T-1} \geq \frac{T^3}{3}$ we have

$$\frac{\mu}{2LS_T} \sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu a^3}{8S_T} a_0 + \frac{4T(T+2a)}{\mu S_T} A + \frac{64T}{\mu^2 S_T} B \quad (73)$$

From convexity we can finally write

$$\mathbb{E} f(\bar{\mathbf{x}}_T) - f^* \leq \frac{La^3}{4S_T} a_0 + \frac{8LT(T+2a)}{\mu^2 S_T} A + \frac{128LT}{\mu^3 S_T} B \quad (74)$$

Where $\bar{\mathbf{x}}_T := \frac{1}{S_T} \sum_{t=0}^{T-1} \left[w_t \left(\frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)} \right) \right] = \frac{1}{S_T} \sum_{t=0}^{T-1} w_t \hat{\mathbf{x}}_t$. This completes the proof of Theorem 3. \square

C Omitted Details from Section 4

As before, in order to prove our results in the asynchronous setting, we define virtual sequences for every worker $r \in [R]$ and for all $t \geq 0$ as follows:

$$\tilde{\mathbf{x}}_0^{(r)} := \hat{\mathbf{x}}_0^{(r)} \quad \tilde{\mathbf{x}}_{t+1}^{(r)} := \tilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}} \left(\hat{\mathbf{x}}_t^{(r)} \right)$$

Define

1. $\tilde{\mathbf{x}}_{t+1} := \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_{t+1}^{(r)} = \tilde{\mathbf{x}}_t - \frac{\eta_t}{R} \sum_{r=1}^R \nabla f_{i_t^{(r)}} \left(\hat{\mathbf{x}}_t^{(r)} \right)$
2. $\mathbf{p}_t := \frac{1}{R} \sum_{r=1}^R \nabla f_{i_t^{(r)}} \left(\hat{\mathbf{x}}_t^{(r)} \right)$
3. $\bar{\mathbf{p}}_t := \mathbb{E}_{(i_t)}[\mathbf{p}_t] = \frac{1}{R} \sum_{r=1}^R \nabla f^{(r)} \left(\hat{\mathbf{x}}_t^{(r)} \right)$
4. $\hat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{x}}_t^{(r)}$
5. $\mathcal{I}_T^{(r)} = \{t_{(i)}^{(r)} : i \in \mathbb{Z}^+, t_{(i)}^{(r)} \in [T], |t_{(i)}^{(r)} - t_{(j)}^{(r)}| \leq H, \forall |i - j| \leq 1\}$

C.1 Proof of Lemma 9

Lemma (Restating Lemma 9). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. For $\widehat{\mathbf{x}}_t^{(r)}$ generated according to Algorithm 2 with decaying learning rate η_t and letting $\widehat{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$, we have the following bound on the deviation of the local sequences:*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq 8(1 + C'' H^2) \eta_t^2 G^2 H^2,$$

where $C'' = 8(4 - 2\gamma)(1 + \frac{C}{\gamma^2})$ and C is a constant satisfying $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$.

Proof. Fix a time t and consider any worker $r \in [R]$. Let $t_r \in \mathcal{I}_T^{(r)}$ denote the last synchronization step until time t for the r 'th worker. Define $t'_0 := \min_{r \in [R]} t_r$. We need to upper-bound $\frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2$. Note that for any R vectors $\mathbf{u}_1, \dots, \mathbf{u}_R$, if we let $\bar{\mathbf{u}} = \frac{1}{R} \sum_{i=1}^R \mathbf{u}_i$, then $\sum_{i=1}^R \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 \leq \sum_{i=1}^R \|\mathbf{u}_i\|^2$. We use this in the first inequality below.

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &= \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \bar{\mathbf{x}}_{t'_0} - (\widehat{\mathbf{x}}_t - \bar{\mathbf{x}}_{t'_0})\|^2 \\ &\leq \frac{1}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \\ &\leq \frac{2}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 + \frac{2}{R} \sum_{r=1}^R \mathbb{E} \|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \end{aligned} \quad (75)$$

We bound both the terms separately. For the first term:

$$\begin{aligned} \mathbb{E} \|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 &= \mathbb{E} \left\| \sum_{j=t_r}^{t-1} \eta_j \nabla f_{i_j^{(r)}}(\widehat{\mathbf{x}}_j^{(r)}) \right\|^2 \\ &\leq (t - t_r) \sum_{j=t_r}^{t-1} \mathbb{E} \|\eta_j \nabla f_{i_j^{(r)}}(\widehat{\mathbf{x}}_j^{(r)})\|^2 \\ &\leq (t - t_r)^2 \eta_{t_r}^2 G^2 \\ &\leq 4\eta_t^2 H^2 G^2. \end{aligned} \quad (76)$$

The last inequality (76) uses $\eta_{t_r} \leq 2\eta_{t_r+H} \leq 2\eta_t$ and $t - t_r \leq H$. To bound the second term of (75), note that we have

$$\bar{\mathbf{x}}_{t_r}^{(r)} = \bar{\mathbf{x}}_{t'_0} - \frac{1}{R} \sum_{s=1}^R \sum_{j=t'_0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(s)}\} g_j^{(s)}. \quad (77)$$

Note that $\widehat{\mathbf{x}}_{t_r}^{(r)} = \bar{\mathbf{x}}_{t_r}^{(r)}$, because at synchronization steps, the local parameter vector becomes equal to the global parameter vector. Using this, the Jensen's inequality, and that $\|\mathbb{1}\{j+1 \in \mathcal{I}_T^{(s)}\} g_j^{(s)}\|^2 \leq \|g_j^{(s)}\|^2$, we can upper-bound (77) as

$$\mathbb{E} \|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \leq \frac{(t_r - t'_0)}{R} \sum_{s=1}^R \sum_{j=t'_0}^{t_r} \mathbb{E} \|g_j^{(s)}\|^2 \quad (78)$$

Now we bound $\mathbb{E}\|g_j^{(s)}\|^2$ for any $j \in \{t'_0, \dots, t_r\}$ and $s \in [R]$: Since $\mathbb{E}\|QComp_k(\mathbf{u})\|^2 \leq B\|\mathbf{u}\|^2$ holds for every \mathbf{u} , with $B = (4 - 2\gamma)$,⁶ we have for any $s \in [R]$ that

$$\mathbb{E}\|g_j^{(s)}\|^2 \leq B\mathbb{E}\|m_j^{(s)} + \mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \quad (79)$$

$$\leq 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \quad (80)$$

Observe that the proof of [Lemma 4](#) does not depend on the synchrony of the network; it only uses the fact that $\text{gap}(\mathcal{I}_T^{(s)}) \leq H$ for any worker $s \in [R]$. Therefore, we can directly use [Lemma 4](#) to bound the first term in (76) as $\mathbb{E}\|m_j^{(s)}\|^2 \leq 4C\frac{\eta_j^2}{\gamma^2}H^2G^2$. In order to bound the second term of (76), note that $\mathbf{x}_j^{(s)} = \widehat{\mathbf{x}}_{t_s}^{(s)}$, which implies that $\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 = \|\sum_{l=t_s}^j \eta_l \nabla f_{i_l}^{(s)}(\widehat{\mathbf{x}}_l^{(s)})\|^2$. Taking expectation yields $\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \leq 4\eta_{t_s}^2 H^2 G^2 \leq 4\eta_{t'_0}^2 H^2 G^2$, where in the last inequality we used that $t'_0 \leq t_s$. Using these in (80) gives

$$\mathbb{E}\|g_j^{(s)}\|^2 \leq 8B \left(1 + \frac{C}{\gamma^2}\right) \eta_{t'_0}^2 H^2 G^2. \quad (81)$$

Since $t'_0 \leq t \leq t'_0 + H$, we have $\eta_{t'_0} \leq 2\eta_{t'_0+H} \leq 2\eta_t$. Putting the bound on $\mathbb{E}\|g_j^{(s)}\|^2$ (after substituting $\eta_{t'_0} \leq 2\eta_t$ in (81)) in (78) gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \leq 32B \left(1 + \frac{C}{\gamma^2}\right) \eta_t^2 H^4 G^2. \quad (82)$$

Putting this and the bound from (76) back in (75) gives

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &\leq 8\eta_t^2 H^2 G^2 + 64B \left(1 + \frac{C}{\gamma^2}\right) \eta_t^2 H^4 G^2 \\ &\leq 8 \left[1 + 8BH^2 \left(1 + \frac{C}{\gamma^2}\right)\right] \eta_t^2 H^2 G^2. \end{aligned}$$

This completes the proof of [Lemma 9](#). □

C.2 Proof of [Lemma 10](#)

Lemma (Restating [Lemma 10](#)). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. By running [Algorithm 2](#) with fixed learning rate η , we have*

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq (2 + H^2 C') \eta^2 G^2 H^2,$$

where $C' = (\frac{16}{\gamma^2} - 12)(4 - 2\gamma)$.

Proof. From (79) and (80) and using the fact that $\mathbb{E}\|QComp_k(\mathbf{u})\|^2 \leq B\|\mathbf{u}\|^2$ for every \mathbf{u} , where $B = (4 - 2\gamma)$, we have the following:

$$\begin{aligned} \mathbb{E}\|g_j^{(s)}\|^2 &\leq 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\eta^2 H^2 G^2 \\ &\leq 8B \frac{(1 - \gamma^2)\eta^2}{\gamma^2} H^2 G^2 + 2\eta^2 B H^2 G^2 \\ &= 2B \left(\frac{4}{\gamma^2} - 3\right) \eta^2 H^2 G^2 \quad (83) \end{aligned}$$

⁶This can be seen as follows: $\mathbb{E}\|QC(\mathbf{u})\|^2 \leq 2\mathbb{E}\|\mathbf{u} - QC(\mathbf{u})\|^2 + 2\|\mathbf{u}\|^2 \leq 2(1 - \gamma)\|\mathbf{u}\|^2 + 2\|\mathbf{u}\|^2$.

For a fixed learning rate η , using (83) and following similar analysis as in (76) we can bound the first term in (75) as follows

$$\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \leq \eta^2 H^2 G^2 \quad (84)$$

Similarly as in (77)-(81) we can bound the second term in (75) as follows

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \leq 2B \left(\frac{4}{\gamma^2} - 3 \right) \eta^2 H^4 G^2 \quad (85)$$

Using (84) and (85) in (75) we can show that

$$\frac{1}{R} \sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \left[2 + 4BH^2 \left(\frac{4}{\gamma^2} - 3 \right) \right] \eta^2 H^2 G^2 \quad (86)$$

This completes the proof of Lemma 10. \square

C.3 Proof of Lemma 11

Lemma (Restating Lemma 11). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. If we run Algorithm 2 with a decaying learning rate η_t , then we have the following bound on the difference between the true and virtual sequences:*

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 \leq C' \eta_t^2 H^4 G^2 + 12C \frac{\eta_t^2}{\gamma^2} G^2 H^2,$$

where $C' = 192(4 - 2\gamma) \left(1 + \frac{C}{\gamma^2} \right)$ and C is a constant satisfying $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$.

Proof. Fix a time t and consider any worker $r \in [R]$. Let $t_r \in \mathcal{I}_T^{(r)}$ denote the last synchronization step until time t for the r 'th worker. Define $t'_0 := \min_{r \in [R]} t_r$. We want to bound $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2$. Note that in the synchronous case, we have shown in Lemma 6 that $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R m_t^{(r)}$. This does not hold in the asynchronous setting, which makes upper-bounding $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2$ a bit more involved. By definition $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \left(\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t^{(r)} \right)$. By the definition of virtual sequences and the update rule for $\widehat{\mathbf{x}}_t^{(r)}$, we also have $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R} \sum_{r=1}^R \left(\widehat{\mathbf{x}}_{t_r}^{(r)} - \widetilde{\mathbf{x}}_{t_r}^{(r)} \right)$. This can be written as

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \left[\frac{1}{R} \sum_{r=1}^R \widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0} \right] + \left[\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t \right] + \left[\bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)} \right] \quad (87)$$

Applying Jensen's inequality and taking expectation gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \left[\frac{3}{R} \sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \right] + \left[3\mathbb{E}\|\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t\|^2 \right] + \left[3\mathbb{E}\|\bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)}\|^2 \right] \quad (88)$$

We bound each of the three terms of (88) separately. We have upper-bounded the first term earlier in (82), which is

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\mathbf{x}}_{t'_0}\|^2 \leq 32B \left(1 + \frac{C}{\gamma^2} \right) \eta_t^2 H^4 G^2, \quad (89)$$

where $B = (4 - 2\gamma)$. To bound the second term of (88), note that

$$\bar{\mathbf{x}}_t = \bar{\mathbf{x}}_0 - \frac{1}{R} \sum_{r=1}^R \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \quad (90)$$

$$= \bar{\mathbf{x}}_{t'_0} - \frac{1}{R} \sum_{r=1}^R \sum_{j=t'_0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \quad (91)$$

By applying Jensen's inequality, using $\|\mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)}\|^2 \leq \|g_j^{(r)}\|^2$, and taking expectation, we can upper-bound (91) as

$$\mathbb{E} \|\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t\|^2 \leq \frac{(t_r - t'_0)}{R} \sum_{r=1}^R \sum_{j=t'_0}^{t_r} \mathbb{E} \|g_j^{(r)}\|^2$$

Using the bound on $\mathbb{E} \|g_j^{(r)}\|^2$'s from (82) gives

$$\mathbb{E} \|\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t\|^2 \leq 32B \left(1 + \frac{C}{\gamma^2}\right) \eta_t^2 H^4 G^2. \quad (92)$$

To bound the last term of (88), note that

$$\tilde{\mathbf{x}}_{t_r}^{(r)} = \bar{\mathbf{x}}_0 - \sum_{j=0}^{t_r-1} \eta_j \nabla f_{i_j}^{(r)} \left(\hat{\mathbf{x}}_j^{(r)}\right) \quad (93)$$

From (90) and (93), we can write

$$\bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_{t_r}^{(r)} = \frac{1}{R} \sum_{r=1}^R \left[\sum_{j=0}^{t_r-1} \eta_j \nabla^{(r)} f_{i_j} \left(\hat{\mathbf{x}}_j^{(r)}\right) - \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \right] \quad (94)$$

Let $t_r^{(1)}$ and $t_r^{(2)}$ be two consecutive synchronization steps in $\mathcal{I}_T^{(r)}$. Then, by the update rule of $\hat{\mathbf{x}}_t^{(r)}$, we have $\hat{\mathbf{x}}_{t_r^{(1)}}^{(r)} - \hat{\mathbf{x}}_{t_r^{(2)} - \frac{1}{2}}^{(r)} = \sum_{j=t_r^{(1)}}^{t_r^{(2)}-1} \nabla f_{i_j}^{(r)} \left(\hat{\mathbf{x}}_j^{(r)}\right)$. Since $\mathbf{x}_{t_r^{(1)}}^{(r)} = \hat{\mathbf{x}}_{t_r^{(1)}}^{(r)}$ and the workers do not modify their local $\mathbf{x}_t^{(r)}$'s in between the synchronization steps, we have $\mathbf{x}_{t_r^{(2)}-1}^{(r)} = \mathbf{x}_{t_r^{(1)}}^{(r)} = \hat{\mathbf{x}}_{t_r^{(1)}}^{(r)}$. Therefore, we can write

$$\mathbf{x}_{t_r^{(2)}-1}^{(r)} - \hat{\mathbf{x}}_{t_r^{(2)} - \frac{1}{2}}^{(r)} = \sum_{j=t_r^{(1)}}^{t_r^{(2)}-1} \nabla f_{i_j}^{(r)} \left(\hat{\mathbf{x}}_j^{(r)}\right). \quad (95)$$

Using (95) for every consecutive synchronization steps, we can equivalently write (94) as

$$\begin{aligned} \bar{\mathbf{x}}_t - \frac{1}{R} \sum_{r=1}^R \tilde{\mathbf{x}}_{t_r}^{(r)} &= \frac{1}{R} \sum_{r=1}^R \left[\sum_{\substack{j: j+1 \in \mathcal{I}_T^{(r)} \\ j \leq t_r-1}} \left(\mathbf{x}_j^{(r)} - \hat{\mathbf{x}}_{j+\frac{1}{2}}^{(r)} - g_j^{(r)} \right) \right] \\ &= \frac{1}{R} \sum_{r=1}^R m_{t_r}^{(r)} \\ &= \frac{1}{R} \sum_{r=1}^R m_t^{(r)} \end{aligned} \quad (96)$$

In the last inequality, we used the fact that the workers do not update their local memory in between the synchronization steps. For the reasons given in the proof of [Lemma 9](#), we can directly apply [Lemma 4](#) to bound the local memories and obtain $\mathbb{E}\|\frac{1}{R}\sum_{r=1}^R m_t^{(r)}\|^2 \leq \frac{1}{R}\sum_{r=1}^R \mathbb{E}\|m_t^{(r)}\|^2 \leq 4C\frac{\eta_t^2}{\gamma^2}G^2H^2$. This implies

$$\mathbb{E}\|\bar{\mathbf{x}}_t - \frac{1}{R}\sum_{r=1}^R \tilde{\mathbf{x}}_{t_r}^{(r)}\|^2 \leq 4C\frac{\eta_t^2}{\gamma^2}G^2H^2. \quad (97)$$

Putting the bounds from [\(89\)](#), [\(92\)](#), and [\(97\)](#) in [\(88\)](#) and using $B = (4 - 2\gamma)$ give

$$\mathbb{E}\|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 \leq 192(4 - 2\gamma) \left(1 + \frac{C}{\gamma^2}\right) \eta_t^2 H^4 G^2 + 12C\frac{\eta_t^2}{\gamma^2}G^2H^2.$$

This completes the proof of [Lemma 11](#). □

C.4 Proof of [Lemma 12](#)

Lemma (Restating [Lemma 12](#)). *Let $\text{gap}(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$. If we run [Algorithm 2](#) with a fixed learning rate η , we have*

$$\mathbb{E}\|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|_2^2 \leq 6C'\eta^2H^4G^2 + \frac{12\eta^2(1 - \gamma^2)}{\gamma^2}G^2H^2,$$

where $C' = (4 - 2\gamma) \left(\frac{8}{\gamma^2} - 6\right)$.

Proof. For a constant learning rate the first term in [\(88\)](#) has been bounded earlier in [\(85\)](#). Following similar steps as in [\(91\)](#) we would have

$$\mathbb{E}\|\bar{\mathbf{x}}_{t'_0} - \bar{\mathbf{x}}_t\|^2 \leq 2B \left(\frac{4}{\gamma^2} - 3\right) \eta^2 H^4 G^2. \quad (98)$$

Finally, using [\(85\)](#),[\(96\)](#), [Lemma 5](#) and [\(98\)](#) in [\(88\)](#) we have

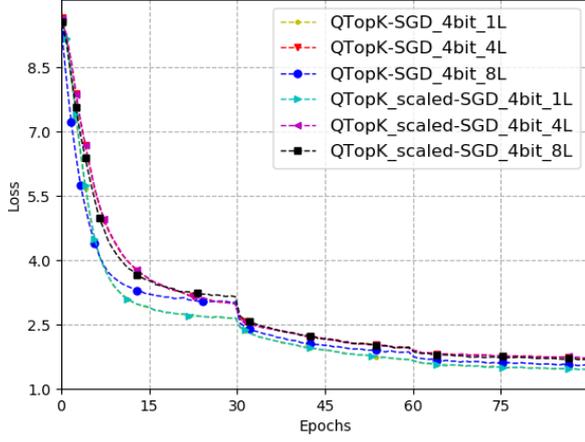
$$\mathbb{E}\|\hat{\mathbf{x}}_t - \tilde{\mathbf{x}}_t\|^2 \leq 12B \left(\frac{4}{\gamma^2} - 3\right) \eta^2 H^4 G^2 + \frac{12\eta^2(1 - \gamma^2)}{\gamma^2}G^2H^2, \quad (99)$$

where $B = (4 - 2\gamma)$. This completes the proof of [Lemma 12](#). □

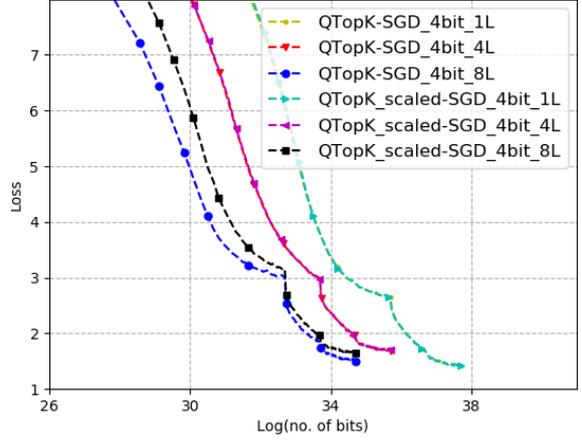
D Omitted Details from [Section 5](#)

As mentioned in [Footnote 4](#), here we compare the performance of Qsparse-local-SGD with scaled and unscaled composed operator $QTop_k$ in the non-convex setting. We will see that even though the scaled $QTop_k$ from [Lemma 2](#) works better than unscaled $QTop_k$ from [Lemma 1](#) theoretically (see [Remark 2](#)), our experiments show the opposite phenomena, that the unscaled $QTop_k$ works at least as good as the scaled $QTop_k$, and strictly better in some cases. We can attribute this to the fact that scaling the composed operator is a sufficient condition to obtain better convergence results, which does not necessarily mean that in practice also it does better. Therefore, we perform our experiments in the non-convex setting [Section 5](#) with unscaled $QTop_k$.

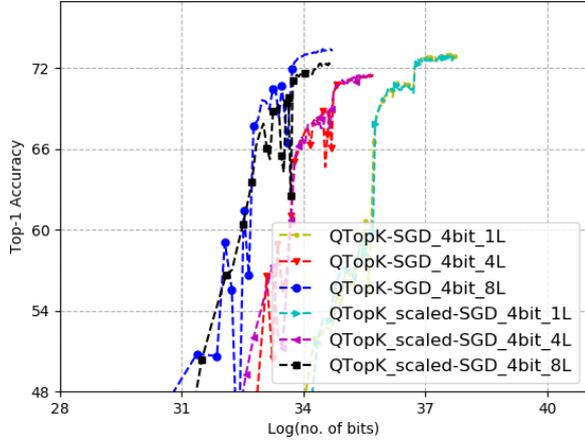
We give plots for the above-mentioned comparison in [Figure 8](#). From [\[AGL⁺17\]](#), we know that for quantized SGD, without any form of error compensation, the dominating term in the convergence rate is affected by the variance blow-up induced due to stochastic quantization;



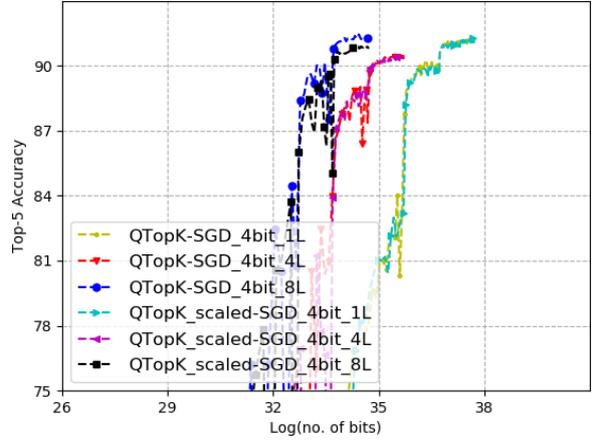
(a) Training loss vs epochs



(b) Training loss vs \log_2 of communication budget



(c) top-1 accuracy [LHS15] for schemes in Figure 8a



(d) top-5 accuracy [LHS15] for schemes in Figure 8a

Figure 8 Figure 8a-8d demonstrate the comparable performance of Q sparse-local-SGD in the non-convex setting with scaled and unscaled $QTop_k$ operators from Lemma 1 and Lemma 2, respectively.

however, with error compensation, we recover rates matching vanilla SGD despite compression and infrequent communication Section 3.3. In Figure 8, $QTop_k$ refers to QSGD composed with the Top_k operator as in Lemma 1, and when used with the subscript *scaled*, we introduce a scaling factor of $(1 + \beta_{k,s})$ as in Lemma 2. Let L denote the number of local iterations in between two synchronization indices. Observe that to achieve a certain target loss or accuracy, both the composed operators perform almost equally in terms of the number of bits transmitted when $L = 0, 4$, but unscaled operator performs better when $L = 8$. Therefore, we restrict our use of composed operator in the non-convex setting to the unscaled $QTop_k$ from Lemma 1.

References

- [ABC⁺16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283, 2016.

- [AGL⁺17] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: communication-efficient SGD via gradient quantization and encoding. In *NIPS*, pages 1707–1718, 2017.
- [AH17] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *EMNLP*, pages 440–445, 2017.
- [AHJ⁺18] D. Alistarh, T. Hoeffler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The convergence of sparsified gradient methods. In *NeurIPS*, pages 5977–5987, 2018.
- [BM11] Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS*, pages 451–459, 2011.
- [Bot10] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186, 2010.
- [BWAA18] J. Bernstein, Y. Wang, K. Azizzadenesheli, and A. Anandkumar. SignSGD: compressed optimisation for non-convex problems. In *ICML*, pages 559–568, 2018.
- [CH16] Kai Chen and Qiang Huo. Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering. In *ICASSP*, pages 5880–5884, 2016.
- [Cop15] Gregory F. Coppola. *Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing*. PhD thesis, University of Edinburgh, UK, 2015.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [GMT73] R. Gitlin, J. Mazo, and M. Taylor. On the design of gradient algorithms for digitally implemented adaptive filters. *IEEE Transactions on Circuit Theory*, 20(2):125–136, March 1973.
- [HK14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.
- [HM51] Robbins Herbert and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*. *JSTOR*, 22, no. 3:400–407, 1951.
- [HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [Kon17] Jakub Konečný. Stochastic, distributed and federated optimization for machine learning. *CoRR*, abs/1707.01155, 2017.
- [KRSJ19] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *ICML*, pages 3252–3261, 2019.

- [KSJ19] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML*, pages 3478–3487, 2019.
- [LBBH98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [LHM⁺18] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *ICLR*, 2018.
- [LHS15] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass SVM. In *NIPS*, pages 325–333, 2015.
- [MMR⁺17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AIS-TATS*, pages 1273–1282, 2017.
- [MPP⁺17] H. Mania, X. Pan, D. S. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.
- [NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [NNvD⁺18] Lam M. Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. SGD and hogwild! convergence without the bounded gradients assumption. In *ICML*, pages 3747–3755, 2018.
- [RB93] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591 vol.1, March 1993.
- [RRWN11] Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pages 693–701, 2011.
- [RSS12] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.
- [SB18] A. Sergeev and M. D. Balso. Horovod: fast and easy distributed deep learning in tensorflow. *CoRR*, abs/1802.05799, 2018.
- [SCJ18] S. U. Stich, J. B. Cordonnier, and M. Jaggi. Sparsified SGD with memory. In *NeurIPS*, pages 4452–4463, 2018.
- [SFD⁺14] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *INTER-SPEECH*, pages 1058–1062, 2014.
- [SSS07] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *ICML*, pages 807–814, 2007.
- [Sti19] Sebastian U. Stich. Local SGD converges fast and communicates little. In *ICLR*, 2019.

- [Str15] Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *INTERSPEECH*, pages 1488–1492, 2015.
- [SYKM17] A. Theertha Suresh, F. X. Yu, S. Kumar, and H. B. McMahan. Distributed mean estimation with limited communication. In *ICML*, pages 3329–3337, 2017.
- [TH12] T. Tieleman and G Hinton. *RMSprop. Coursera: Neural Networks for Machine Learning, Lecture 6.5*. 2012.
- [WHHZ18] J. Wu, W. Huang, J. Huang, and T. Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *ICML*, pages 5321–5329, 2018.
- [WJ18] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *CoRR*, abs/1808.07576, 2018.
- [WSL⁺18] H. Wang, S. Sievert, S. Liu, Z. B. Charles, D. S. Papailiopoulos, and S. Wright. ATOMO: communication-efficient learning via atomic sparsification. In *NeurIPS*, pages 9872–9883, 2018.
- [WWLZ18] J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. In *NeurIPS*, pages 1306–1316, 2018.
- [WXY⁺17] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *NIPS*, pages 1508–1518, 2017.
- [WYL⁺18] Tianyu Wu, Kun Yuan, Qing Ling, Wotao Yin, and Ali H. Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Trans. Signal and Information Processing over Networks*, 4(2):293–307, 2018.
- [YJY19] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *ICML*, pages 7184–7193, 2019.
- [YYZ19] 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 *AAAI*, pages 5693–5700, 2019.
- [ZDJW13] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, pages 2328–2336, 2013.
- [ZDW13] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.
- [ZSMR16] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel SGD: when does averaging help? *CoRR*, abs/1606.07365, 2016.