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

Debraj Basu *
Adobe Inc.
dbasu@adobe.com

Deepesh Data
UCLA
deepeshdata@ucla.edu

Can Karakus *
Amazon Inc.
cakarak@amazon.com

Suhas Diggavi
UCLA
suhasdiggavi@ucla.edu

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* case, 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.

1 Introduction

Stochastic Gradient Descent (SGD) [14] and its many variants have become the workhorse for modern large-scale optimization as applied to machine learning [5, 8]. We consider the setup where SGD is applied to the *distributed* setting, where R different nodes compute *local* SGD 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 $\{g_t^r\}_{r=1}^R$ are the local stochastic gradients at the R machines for a local loss function $f^{(r)}(\mathbf{x})$ of the parameters, where $f^{(r)} : \mathbb{R}^d \to \mathbb{R}$.

It is well understood by now that sending full-precision gradients, causes communication to be the bottleneck for many large scale models [4, 7, 33, 39]. The communication bottleneck could be significant in emerging edge computation architectures suggested by federated learning [1, 17, 22]. To address this, many methods have been proposed recently, and these methods are broadly based on three major approaches: (i) *Quantization* of gradients, where nodes locally quantize the gradient (perhaps with randomization) to a small number of bits [3,7,33,39,40]. (ii) *Sparsification* of gradients, *e.g.,* where nodes locally select Top_k values of the gradient in absolute value and transmit these at full precision [2, 4, 20, 30, 32, 40], while maintaining errors in local nodes for later compensation. (iii) *Skipping communication rounds* whereby nodes average their models after locally updating their models for several steps [9, 10, 31, 34, 37, 43, 45].

*Work done while Debraj Basu and Can Karakus were at UCLA.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

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[2] implementations of *Qsparse-local-SGD*. We analyze convergence for *Qsparse-local-SGD* in the *distributed* case, 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 implement *Qsparse-local-SGD* for ResNet-50 using the ImageNet dataset, and show that we achieve target accuracies with a small penalty in final accuracy (approximately 1 %), with about a factor of 15-20 savings over the state-of-the-art [4, 30, 31], in the total number of bits transmitted. While the downlink communication is not our focus in this paper (also in [4, 20, 39], for example), it can be inexpensive when the broadcast routine is implemented in a tree-structured manner as in many MPI implementations, or if the parameter server aggregates the sparse quantized updates and broadcasts it.

Related work. The use of quantization for communication efficient gradient methods has decades rich history [11] and its recent use in training deep neural networks [27, 32] has re-ignited interest. Theoretically justified gradient compression using unbiased stochastic quantizers has been proposed and analyzed in [3, 33, 39]. Though methods in [36, 38] 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 [2, 4, 20, 30, 32], combined with error compensation to ensure that all co-ordinates do get eventually updated as needed. [40] analyzed error compensation for QSGD, without Top_k sparsification and a focus 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* and *model averaging*, see [31, 43] and references therein. Our work is most closely related to and builds on the recent theoretical results in [4, 30, 31, 43]. [30] considered the analysis for the centralized Top_k (among other sparsifiers), and [4] 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 [4]. [31, 43] studied local-SGD, where several local iterations are done before sending the *full* gradients, and did not do 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 [4], by using the perturbed iterate methods [21, 30]. We analyze non-convex (smooth) objectives as well as strongly convex objectives for the distributed case with local computations. [30] gave a proof only for convex objective functions and for centralized case and therefore without local computations[3]. Our techniques compose a (stochastic or deterministic 1-bit sign) quantizer with sparsification and local computations using error compensation; in fact this technique works for any compression operator satisfying a regularity condition (see Definition 3).

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 $f^{(r)}(\mathbf{x}) = \mathop{\mathbb{E}}_{i\sim\mathcal{D}_r}[f_i(\mathbf{x})]$, where $\mathop{\mathbb{E}}_{i\sim\mathcal{D}_r}[\cdot]$ denotes expectation[4] over a random sample chosen from the local data set $\mathcal{D}_r$. For $f : \mathbb{R}^d \to \mathbb{R}$, 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.

[2] 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.

[3] At the completion of our work, we recently found that in parallel to our work [15] examined use of sign-SGD quantization, *without sparsification* for the centralized model. Another recent work in [16] studies the decentralized case with sparsification for strongly convex function. Our work, developed independent of these works, uses quantization, sparsification and local computations for the distributed case with local computations for both non-convex and strongly convex objectives.

[4] 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.

Our main theoretical results are the convergence analysis of *Qsparse-local-SGD* for both (smooth) non-convex objectives as well as for the strongly convex case. See Theorem 1, 2 for the synchronous case, as well as Theorem 3, 4, for the asynchronous operation. Our analysis also demonstrates natural gains in convergence that distributed, mini-batch operation affords, and has convergence similar to vanilla SGD with local iterations (see Corollary 1, 2), for both the non-convex case (with convergence rate $\sim {}^1/\sqrt{T}$ for fixed learning rate) as well as the strongly convex case (with convergence rate $\sim {}^1/T$, for diminishing learning rate), demonstrating that quantizing and sparsifying the gradient, even after local iterations asymptotically yields an almost "free" communication efficiency gain (also observed numerically in Section 5 non-asymptotically). The numerical results on ImageNet dataset implemented for a ResNet-50 architecture demonstrates that one can get significant communication savings, while retaining equivalent state-of-the art performance with a small penalty in final accuracy.

Unlike previous works, *Qsparse-local-SGD* stores the compression error of the net *local update*, which is a sum of at most H gradient steps and the historical error, in the local memory. From literature [4, 30], we know that methods with error compensation work only when the evolution of the error is controlled. The combination of quantization, sparsification, and local computations poses several challenges for theoretical analysis, including (i) the analysis of impact of local iterations on the evolution of the error due to quantization and sparsification, as well as the deviation of local iterates (see Lemma 3, 4, 8, 9) (ii) asynchronous updates together with distribution compression using operators which satisfy Definition 3, including our composed (*Qsparse*) operators. (see Lemma 11-14 in appendix). Another useful technical observation is that the composition of a quantizer and a sparsifier results in a compression operator (Lemma 1, 2); see Appendix A for proofs on the same.

We provide additional results in the appendices as part of the supplementary material. These include results on the asymptotic analysis for non-convex objectives in Theorem 5, 8 along with precise statements of the convergence guarantees for the asynchronous operation Theorem 6, 7 and numerics for the convex case for multi-class logistic classification on *MNIST* [19] dataset in Appendix D, for both synchronous and asynchronous operations.

We believe that our approach for combining different forms of compression and local computations can be extended to the decentralized case, where nodes are connected over an arbitrary graph, building on the ideas from [15, 35]. Our numerics also incorporate momentum acceleration, whose analysis is a topic for future research, for example incorporating ideas from [42].

Organization. In Section 2, we demonstrate that composing certain classes of quantization with sparsification 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 3.1, briefly giving the proof ideas in Section 3.2. 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 and additional results are given in the appendices provided with the supplementary material.

2 Composition of Quantization and Sparsification

In this section, we consider composition of 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 [7, 15] or randomized [3, 33, 39, 44] 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 or choosing k components uniformly at random, denoted by Rand_k, [30].

Definition 1 (Randomized Quantizer [3, 33, 39, 44])**.** *We say that $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ 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) $\mathbb{E}_Q[\|Q_s(\mathbf{x})\|^2] \le (1 + \beta_{d,s})\|\mathbf{x}\|^2$*, 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 (i) *QSGD* [3, 39], which independently quantizes components of $\mathbf{x} \in \mathbb{R}^d$ into s levels, with $\beta_{d,s} = \min(\frac{d}{s^2}, \frac{\sqrt{d}}{s})$; (ii) *Stochastic s-level Quantization* [33, 44], which independently quantizes every component of $\mathbf{x} \in \mathbb{R}^d$ into s levels between $\text{argmax}_i x_i$ and $\text{argmin}_i x_i$, with $\beta_{d,s} = \frac{d}{2s^2}$; and (iii) *Stochastic Rotated Quantization* [33], 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 to the nearest level. In particular, we can just take the sign, which has shown promise in [7, 27, 32].

Definition 2 (Deterministic Sign Quantizer [7, 15])**.** *A deterministic quantizer* $Sign : \mathbb{R}^d \to \{+1, -1\}^d$ *is defined as follows: for every vector* $\mathbf{x} \in \mathbb{R}^d$*,* $i \in [d]$*, the* i*'th component of* $Sign(\mathbf{x})$ *is defined as* $\mathbb{1}\{x_i \geq 0\} - \mathbb{1}\{x_i < 0\}$*.*

As mentioned above, we consider two important examples of sparsification operators: Top_k and Rand_k, For any $\mathbf{x} \in \mathbb{R}^d$, $\mathrm{Top}_k(\mathbf{x})$ 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, $\mathrm{Rand}_k(\mathbf{x})$ 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 $\gamma = k/d$ [30]. Few other examples of such operators can be found in [30].

Definition 3 (Sparsification [30])**.** *A (randomized) function* $Comp_k : \mathbb{R}^d \to \mathbb{R}^d$ *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* $\mathbb{E}_C[\|\mathbf{x} - Comp_k(\mathbf{x})\|_2^2] \leq (1-\gamma)\|\mathbf{x}\|_2^2$*, where expectation is taken over* $Comp_k$*.*

We can apply different compression operators to different coordinates of a vector, and the resulting operator is also a compression operator; see Corollary 3 in Appendix A. As an application, in the case of training neural networks, we can apply different compression operators to different layers.

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. Proofs are given in Appendix A.

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

Lemma 2 (Composing sparsification with deterministic quantization)**.** *Let* $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$*. Let* $SignComp_k : \mathbb{R}^d \to \mathbb{R}^d$ *be defined as follows: for every* $\mathbf{x} \in \mathbb{R}^d$*, the* i*'th component of* $SignComp_k(\mathbf{x})$ *is equal to* $\mathbb{1}\{x_i \geq 0\} - \mathbb{1}\{x_i < 0\}$*, if the* i*'th component is chosen in defining* $Comp_k$*, otherwise, it is equal to* 0*. Then* $\frac{\|Comp_k(\mathbf{x})\|_1 \, SignComp_k(\mathbf{x})}{k}$ *is a compression operator*[5] *with the compression coefficient being equal to* $\gamma = \max\left\{\frac{1}{d}, \frac{k}{d}\left(\frac{\|Comp_k(\mathbf{x})\|_1}{\sqrt{d}\|Comp_k(\mathbf{x})\|_2}\right)^2\right\}$*.*

3 Qsparse-local-SGD

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 $\widehat{\mathbf{x}}_t^{(r)}$ which is updated in each iteration t, using the stochastic gradient $\nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$, where $i_t^{(r)}$ is a mini-batch of size b sampled uniformly in $\mathcal{D}_r$. If $t \in \mathcal{I}_T$, the sparsified error-compensated update $g_t^{(r)}$ computed on the net progress made since the last synchronization is sent to the master node, and updates its local memory $m_t^{(r)}$. Upon receiving $g_t^{(r)}$'s from every worker, 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.

3.1 Main Results for Synchronous Operation

All results in this paper use the following two standard assumptions. (i) **Smoothness:** The local function $f^{(r)} : \mathbb{R}^d \to \mathbb{R}$ at each worker $r \in [R]$ is L-smooth, i.e., for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $f^{(r)}(\mathbf{y}) \leq f^{(r)}(\mathbf{x}) + \langle \nabla f^{(r)}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$. (ii) **Bounded second moment:** For every

[5] The analysis for general p-norm, i.e. $\frac{\|Comp_k(\mathbf{x})\|_p \, SignComp_k(\mathbf{x})}{k}$, for any $p \in \mathbb{Z}_+$ is provided in Appendix A.

Algorithm 1 Qsparse-local-SGD

Initialize $\mathbf{x}_0 = \widehat{\mathbf{x}}_0^{(r)} = m_0^{(r)},\ \forall r \in [R]$. Suppose η_t follows a certain learning rate schedule.
for $t = 0$ **to** $T-1$ **do**
  On Workers:
  for $r = 1$ **to** R **do**
    $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} \leftarrow \widehat{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$; $i_t^{(r)}$ is a mini-batch of size b sampled uniformly in $\mathcal{D}_r$
    if $t+1 \notin \mathcal{I}_T$ **then**
      $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$, $m_{t+1}^{(r)} \leftarrow m_t^{(r)}$ and $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$
    else
      $g_t^{(r)} \leftarrow QComp_k\left(m_t^{(r)} + \mathbf{x}_t - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}\right)$, send $g_t^{(r)}$ to the master.
      $m_{t+1}^{(r)} \leftarrow m_t^{(r)} + \mathbf{x}_t - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} - g_t^{(r)}$
      Receive $\mathbf{x}_{t+1}$ from the master and set $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \mathbf{x}_{t+1}$
    end if
  end for
  At Master:
  if $t+1 \notin \mathcal{I}_T$ **then**
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$
  else
    Receive $g_t^{(r)}$ from R workers and compute $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{R}\sum_{r=1}^{R} g_t^{(r)}$
    Broadcast $\mathbf{x}_{t+1}$ to all workers.
  end if
end for
Comment: Note that $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$ is used to denote an intermediate variable between iterations t and $t+1$.

$\widehat{\mathbf{x}}_t^{(r)} \in \mathbb{R}^d, r \in [R], t \in [T]$, we have $\mathbb{E}_{i\sim\mathcal{D}_r}[\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)})\|^2] \leq G^2$, for some constant $G < \infty$. This is a standard assumption in [4, 12, 16, 23, 25, 26, 29–31, 43]. Relaxation of the uniform boundedness of the gradient allowing arbitrarily different gradients of local functions in heterogenous settings as done for SGD in [24, 37] is left as future work. This also imposes a **bound on the variance**: $\mathbb{E}_{i\sim\mathcal{D}_r}[\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)}) - \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2] \leq \sigma_r^2$, where $\sigma_r^2 \leq G^2$ for every $r \in [R]$. To state our results, we need the following definition from [31].

Definition 4 (Gap [31])**.** *Let $\mathcal{I}_T = \{t_0, t_1, \ldots, t_k\}$, where $t_i < t_{i+1}$ for $i = 0, 1, \ldots, k-1$. The gap of $\mathcal{I}_T$ is defined as $gap(\mathcal{I}_T) := \max_{i\in[k]}\{(t_i - t_{i-1})\}$, which is equal to the maximum difference between any two consecutive synchronization indices.*

We leverage the perturbed iterate analysis as in [21, 30] to provide convergence guarantees for *Qsparse-local-SGD*. Under assumptions (i) and (ii), the following theorems hold when Algorithm 1 is run with any compression operator (including our composed operators).

Theorem 1 (Convergence in the smooth (non-convex) case with fixed learning rate)**.** *Let $f^{(r)}(\mathbf{x})$ be L-smooth for every $i \in [R]$. Let $QComp_k : \mathbb{R}^d \to \mathbb{R}^d$ 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 $gap(\mathcal{I}_T) \leq H$. Then we have*

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

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

Corollary 1. *Let $\mathbb{E}[f(\mathbf{x}_0)] - f^* \leq J^2$, where $J < \infty$ is a constant,*[6] *$\sigma_{max} = \max_{r\in[R]} \sigma_r$, and $\widehat{C}^2 = \frac{bR(\mathbb{E}[f(\mathbf{x}_0)] - f^*)}{\sigma_{max}^2 L}$, we have*

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

[6]Even 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.

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*[7] $H = \mathcal{O}\left(\gamma T^{1/4}/(bR)^{3/4}\right)$.

Theorem 1 is proved in Appendix B and provides non-asymptotic guarantees, where we observe that compression does not affect the first order term. The corresponding asymptotic result (with decaying learning rate), with a convergence rate of $\mathcal{O}(\frac{1}{\log T})$, is provided in Theorem 5 in Appendix B.

Theorem 2 (Convergence in the smooth and strongly convex case with a decaying learning rate)**.** *Let* $f^{(r)}(\mathbf{x})$ *be* L*-smooth and* μ*-strongly convex. Let* $QComp_k : \mathbb{R}^d \to \mathbb{R}^d$ *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* $gap(\mathcal{I}_T) \leq H$*, where* $a > 1$ *is such that we have* $a \geq \max\{{}^{4H}/_{\gamma}, 32\kappa, H\}$, $\kappa = {}^L/_{\mu}$*. Then the following holds*

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

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) $\overline{\mathbf{x}}_T := \frac{1}{S_T}\sum_{t=0}^{T-1}\left[w_t\left(\frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}\right)\right]$, *where* $w_t = (a+t)^2$*; and* (iii) $S_T = \sum_{t=o}^{T-1} w_t \geq \frac{T^3}{3}$.

Corollary 2. *For* $a > \max\{\frac{4H}{\gamma}, 32\kappa, H\}$, $\sigma_{max} = \max_{r\in[R]}\sigma_r$*, and using* $\mathbb{E}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \frac{4G^2}{\mu^2}$ *from Lemma 2 in [25], we have*

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

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 2 has been proved in Appendix B. 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: (i) We recover [43, Theorem 1], which is for non-convex case; (ii) We generalize [31, Theorem 2.2], which is for a strongly convex case and requires that each worker has identical datasets, to the distributed case. We emphasize that unlike [31, 43], which only consider local computation, we combine quantization and sparsification with local computation, which poses several technical challenges (e.g., see proofs of Lemma 3, 4,7 in Appendix B).

3.2 Proof Outlines

Maintain virtual sequences for every worker

$$\widetilde{\mathbf{x}}_0^{(r)} := \widehat{\mathbf{x}}_0^{(r)} \quad \text{and} \quad \widetilde{\mathbf{x}}_{t+1}^{(r)} := \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \tag{5}$$

Define (i) $\mathbf{p}_t := \frac{1}{R}\sum_{r=1}^R \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$, $\overline{\mathbf{p}}_t := \mathbb{E}_{i_t}[\mathbf{p}_t] = \frac{1}{R}\sum_{r=1}^R \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$;

and (ii) $\widetilde{\mathbf{x}}_{t+1} := \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t - \eta_t\mathbf{p}_t$, $\widehat{\mathbf{x}}_t := \frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$.

Proof outline of Theorem 1. Since f is L-smooth, we have $f(\widetilde{\mathbf{x}}_{t+1}) - f(\widetilde{\mathbf{x}}_t) \leq -\eta_t\langle\nabla f(\widetilde{\mathbf{x}}_t), \mathbf{p}_t\rangle + \frac{\eta_t^2 L}{2}\|\mathbf{p}_t\|^2$. With some algebraic manipulations provided in Appendix B, for $\eta_t \leq {}^1/_{2L}$, we arrive at

$$\frac{\eta_t}{4R}\sum_{r=1}^R \mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \eta_t^2 L\mathbb{E}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 + 2\eta_t L^2\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2$$
$$+ 2\eta_t L^2 \frac{1}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2. \tag{6}$$

Under Assumptions 1 and 2, we have $\mathbb{E}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 \leq \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$. To bound $\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2$ in (6), we first show (in Lemma 7 in Appendix B) that $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R m_t^{(r)}$, i.e., the difference of the true and the virtual parameter vectors is equal to the average memory, and then we bound the local memory at each worker $r \in [R]$ below.

[7] Here we characterize the reduction in communication that can be afforded, however for a constant H we get the same rate of convergence after $T = \Omega\left((bR)^3/\gamma^4\right)$. Analogous statements hold for Theorem 2-4.

Lemma 3 (Bounded Memory)**.** *For $\eta_t = \eta$, $gap(\mathcal{I}_T) \leq H$, we have for every $t \in \mathbb{Z}^+$ that*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\tfrac{\eta^2(1-\gamma^2)}{\gamma^2}H^2G^2. \tag{7}$$

Using Lemma 3, we get $\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 \leq \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta^2(1-\gamma^2)}{\gamma^2}H^2G^2$. We can bound the last term of (6) as $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \eta^2 G^2 H^2$ in Lemma 9 in Appendix B. Putting them back in (6), performing a telescopic sum from $t=0$ to $T-1$, and then taking an average over time, we get

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

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 Theorem 1. □

Proof outline of Theorem 2. Using the definition of virtual sequences (5), we have $\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 - 2\eta_t\left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t, \mathbf{p}_t - \overline{\mathbf{p}}_t\right\rangle$. With some algebraic manipulations provided in Appendix B, for $\eta_t \leq 1/4L$ and letting $e_t = \mathbb{E}[f(\widehat{\mathbf{x}}_t)] - f^*$, we get

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \tfrac{\mu\eta_t}{2}\right)\mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\eta_t\mu}{2L}e_t + \eta_t\left(\tfrac{3\mu}{2} + 3L\right)\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad + \tfrac{3\eta_t L}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 + \eta_t^2\tfrac{\sum_{r=1}^{R}\sigma_r^2}{bR^2}.\end{aligned} \tag{8}$$

To bound the 3rd term on the RHS of (63), first we note that $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^{R}m_t^{(r)}$, and then we bound the local memory at each worker $r \in [R]$ below.

Lemma 4 (Memory Contraction)**.** *For $a > {}^{4H}\!/_{\gamma}$, $\eta_t = {}^{\xi}\!/_{a+t}$, $gap(\mathcal{I}_T) \leq H$, there exists a $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$ such that the following holds for every $t \in \mathbb{Z}^+$*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\tfrac{\eta_t^2}{\gamma^2}CH^2G^2. \tag{9}$$

A proof of Lemma 4 is provided in Appendix B and is technically more involved than the proof of Lemma 3. This complication arises because of the decaying learning rate, combined with compression and local computation. We can bound the penultimate term on the RHS of (63) as $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2G^2H^2$. This can be shown along the lines of the proof of [31, Lemma 3.3] and we show it in Lemma 8 in Appendix B. Substituting all these in (63) gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \tfrac{\mu\eta_t}{2}\right)\mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\mu\eta_t}{2L}e_t + \eta_t\left(\tfrac{3\mu}{2} + 3L\right)C\tfrac{4\eta_t^2}{\gamma^2}G^2H^2 \\ &\quad + (3\eta_t L)4\eta_t^2LG^2H^2 + \eta_t^2\tfrac{\sum_{r=1}^{R}\sigma_r^2}{bR^2}.\end{aligned} \tag{10}$$

Since (10) is a contracting recurrence relation, with some calculation done in Appendix B, we complete the proof of Theorem 2. □

4 Asynchronous Qsparse-local-SGD

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 [17, 22]. 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 synchronous with respect to the global clock.[8]

In this asynchronous setting, $\mathcal{I}_T^{(r)}$'s may be different for different workers. However, we assume that $gap(\mathcal{I}_T^{(r)}) \leq H$ 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 in Appendix C. We give the simplified expressions of our main results below; more precise results are in Appendix C.

[8]This is different from asynchronous algorithms studied for stragglers [26, 41], where only one gradient step is taken but occurs at different times due to delays.

Theorem 3 (Convergence in the smooth non-convex case with fixed learning rate)**.** *Under the same conditions as in Theorem 1 with $gap(\mathcal{I}_T^{(r)}) \leq H$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to Algorithm 2, the following holds, where $\mathbb{E}[f(\mathbf{x}_0)] - f^* \leq J^2$, $\sigma_{max} = \max_{r\in[R]} \sigma_r$, and $\widehat{C}^2 = {}^{bR(\mathbb{E}[f(\mathbf{x}_0)]-f^*)}/_{\sigma_{max}^2}$.*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|^2 \leq \mathcal{O}\left(\frac{J\sigma_{max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2 bRG^2}{\sigma_{max}^2\gamma^2 T}(H^2 + H^4)\right). \tag{11}$$

where $\mathbf{z}_T$ is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability $1/RT$. 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 $H = \mathcal{O}\left(\sqrt{\gamma}T^{1/8}/(bR)^{3/8}\right)$.

We give a precise result in Theorem 6 in Appendix C. Note that Theorem 3 provides non-asymptotic guarantees, where compression is almost for "free". The corresponding asymptotic result with decaying learning rate, with a convergence rate of $\mathcal{O}(\frac{1}{\log T})$, is provided in Theorem 8 in Appendix C.

Theorem 4 (Convergence in the smooth and strongly convex case with decaying learning rate)**.** *Under the same conditions as in Theorem 2 with $gap(\mathcal{I}_T^{(r)}) \leq H$, $a > \max\{{}^{4H}/_{\gamma}, 32\kappa, H\}$, $\sigma_{max} = \max_{r\in[R]} \sigma_r$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to Algorithm 2, the following holds:*

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \mathcal{O}\left(\frac{G^2H^3}{\mu^2\gamma^3T^3}\right) + \mathcal{O}\left(\frac{\sigma_{max}^2}{\mu^2 bRT} + \frac{H\sigma_{max}^2}{\mu^2 bR\gamma T^2}\right) + \mathcal{O}\left(\frac{G^2}{\mu^3\gamma^2T^2}(H^2 + H^4)\right). \tag{12}$$

where $\overline{\mathbf{x}}_T$, S_T are as defined in Theorem 2. To ensure that the compression does not affect the dominating terms while converging at a rate of $\mathcal{O}(1/(bRT))$, we would require $H = \mathcal{O}\left(\sqrt{\gamma}(T/(bR))^{1/4}\right)$.

We give a more precise result in Theorem 7 in Appendix C. If $\mathcal{I}_T^{(r)}$'s are the same for all the workers, then one would ideally require that the bounds on H in the asynchronous setting reduce to the bounds on H in the synchronous setting. This is not happening, as our bounds in the asynchronous setting are for the worst case scenario – they hold as long as $gap(\mathcal{I}_T^{(r)}) \leq H$, for every $r \in [R]$.

4.1 Proof Outlines

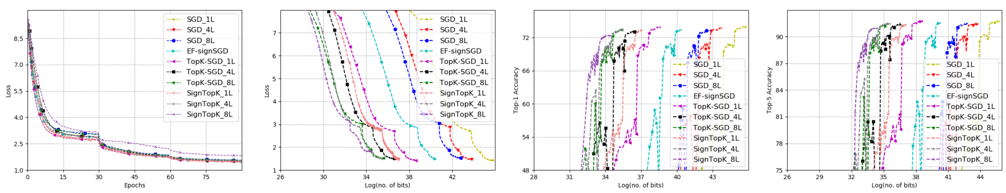
Our proofs of these results follow the same outlines of the corresponding proofs in the synchronous setting, but some technical details change significantly. This is because, in our asynchronous setting, workers are allowed to update the global parameter vector in between two consecutive synchronization time steps of other workers. For example, unlike the synchronous setting, $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^{R} m_t^{(r)}$ does not hold here; however, we can show that $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t$ 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 $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \mathcal{O}\left({}^{\eta_t^2}/_{\gamma^2}G^2(H^2 + H^4)\right)$ (vs. $\mathcal{O}\left({}^{\eta_t^2}/_{\gamma^2}G^2H^2\right)$ for the synchronous setting), proved in Lemma 13-14 in Appendix C. Similarly, the proof of the average true sequence being close to the virtual sequence requires carefully chosen reference points on the global parameter sequence lying within bounded steps of the local parameters. We show a bound on $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \mathcal{O}(\eta_t^2 G^2(H^2 + {}^{H^4}/_{\gamma^2}))$, which is weaker than the corresponding bound $\mathcal{O}(\eta_t^2G^2H^2)$ for the synchronous setting, in Lemma 11-12 in Appendix C.

5 Experiments

Experiment setup: We train ResNet-50 [13] (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 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 [28].[9] We use $SignTop_k$ (as in Lemma 2) as our composed operator. 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. We also perform analogous experiments on the *MNIST* [19] handwritten digits dataset for softmax regression with a standard ℓ_2 regularizer, using the synchronous operation of *Qsparse-local-SGD* with 15 workers, and

[9]Our implementation is available at https://github.com/karakusc/horovod/tree/qsparselocal.

a decaying learning rate as proposed in Theorem 2, the details of which are provided in Appendix D.[10]
Results: Figure 1 compares the performance of $SignTop_k$-SGD (which employs the 1 bit sign quantizer and the Top_k sparsifier) with error compensation (SignTopK) against (i) Top_k SGD with error compensation (TopK-SGD), (ii) SignSGD with error compensation (EF-SIGNSGD), and (iii) vanilla SGD (SGD). All of these are specializations of *Qsparse-local-SGD*. Furthermore, SignTopK_hL uses a synchronization period of h; same applies for other schemes. From Figure 1a, we observe that quantization and sparsification, both individually and combined, with error compensation, has almost no penalty in terms of convergence rate, with respect to vanilla SGD. We observe that SignTopK demonstrates superior performance over EF-SIGNSGD, TopK-SGD, as well as vanilla SGD, both in terms of the required number of communicated bits for achieving a certain target loss as well as test accuracy. This is because in SignTopK, we send only 1 bit for the sign of each Top_k coordinate, along with its location. Observe that the incorporation of local iterations in Figure 1a has very little impact on the convergence rates, as compared to vanilla SGD with the same number of local iterations. Furthermore, this provides an added advantage over SignTopK, in terms of savings (by a factor of 6 to 8 times on average) in communication bits for achieving a certain target loss; see Figure 1b.



(a) Training loss vs epochs

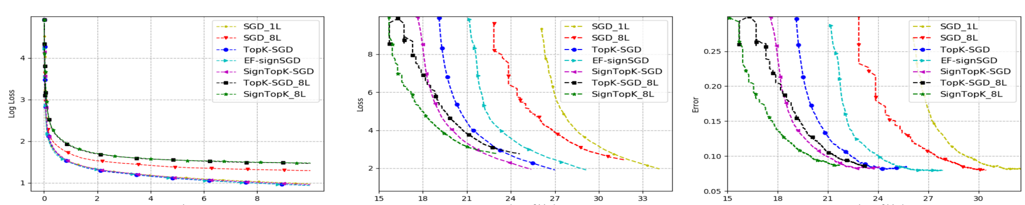
(b) Training loss vs $\log_2$ of communication budget

(c) top-1 accuracy [18] for schemes in Figure 1a

(d) top-5 accuracy [18] for schemes in Figure 1a

Figure 1 Figure 1a-1d demonstrate performance gains of our of our scheme in comparison with local SGD [31], EF-SIGNSGD [15] and TopK-SGD [4, 30] in a non-convex setting for synchronous updates.

Figure 1c and Figure 1d show the top-1, and top-5 convergence rates,[11] respectively, with respect to the total number of bits of communication used. We observe that *Qsparse-local-SGD* combines the bit savings of the deterministic sign based operator and aggressive sparsifier, 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 accuracy in the case of *Qsparse-local-SGD* is around 1/16 in comparison with TopK-SGD and over $1000\times$ less than vanilla SGD.



(a) Training loss vs epochs

(b) Training loss vs $\log_2$ of communication budget

(c) top-1 accuracy [18] for schemes in Figure 2a

Figure 2 Figure 2a-2c demonstrate the performance gains of our scheme in a convex setting.

Figure 2b and 2c makes similar comparisons in the convex setting, and shows that for a test error approximately 0.1, *Qsparse-local-SGD* combines the benefits of the composed operator $SignTop_k$, with local computations, and needs 10-15 times less bits than TopK-SGD and $1000\times$ less bits than vanilla SGD. Also in Figure 2a, we observe that both TopK-SGD and SignTopK_8L (SignTopK with 8 local iterations) converge at rates which are almost similar to that of their corresponding local SGD counterpart. Our experiments in both non-convex and convex settings verify 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.

[10] Further numerics demonstrating the performance of *Qsparse-local-SGD* for the composition of a stochastic quantizer with a sparsifier, as compared to $SignTop_k$ and other standard baselines can be found in [6].

[11] top-i refers to the accuracy of the top i predictions by the model from the list of possible classes; see [18].

Acknowledgments

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.

References

[1] 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.

[2] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *EMNLP*, pages 440–445, 2017.

[3] 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.

[4] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. The convergence of sparsified gradient methods. In *NeurIPS*, pages 5977–5987, 2018.

[5] Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS*, pages 451–459, 2011.

[6] Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-sgd: Distributed SGD with quantization, sparsification, and local computations. *CoRR*, abs/1906.02367, 2019.

[7] J. Bernstein, Y. Wang, K. Azizzadenesheli, and A. Anandkumar. SignSGD: compressed optimisation for non-convex problems. In *ICML*, pages 559–568, 2018.

[8] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186, 2010.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] Robbins Herbert and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics. JSTOR*, 22, no. 3:400–407, 1951.

[15] 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.

[16] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML*, pages 3478–3487, 2019.

[17] Jakub Konecný. Stochastic, distributed and federated optimization for machine learning. *CoRR*, abs/1707.01155, 2017.

[18] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass SVM. In *NIPS*, pages 325–333, 2015.

[19] 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.

[20] 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.

[21] 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.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[23] 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.

[24] 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.

[25] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

[26] 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.

[27] 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 *INTERSPEECH*, pages 1058–1062, 2014.

[28] A. Sergeev and M. D. Balso. Horovod: fast and easy distributed deep learning in tensorflow. *CoRR*, abs/1802.05799, 2018.

[29] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *ICML*, pages 807–814, 2007.

[30] S. U. Stich, J. B. Cordonnier, and M. Jaggi. Sparsified SGD with memory. In *NeurIPS*, pages 4452–4463, 2018.

[31] Sebastian U. Stich. Local SGD converges fast and communicates little. In *ICLR*, 2019.

[32] Nikko Strom. Scalable distributed DNN training using commodity GPU cloud computing. In *INTERSPEECH*, pages 1488–1492, 2015.

[33] A. Theertha Suresh, F. X. Yu, S. Kumar, and H. B. McMahan. Distributed mean estimation with limited communication. In *ICML*, pages 3329–3337, 2017.

[34] H. Tang, S. Gan, C. Zhang, T. Zhang, and Ji Liu. Communication compression for decentralized training. In *NeurIPS*, pages 7663–7673, 2018.

[35] Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *NeurIPS*, pages 7663–7673, 2018.

[36] 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.

[37] 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.

[38] J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. In *NeurIPS*, pages 1306–1316, 2018.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

A Supplementary Material of Section 2

Now we give a simple but important corollary, which allows us to apply different compression operators to different coordinates of a vector. For example, in the case of training neural networks, we can apply different compression operator to different layers.

Corollary 3 (Piecewise Compression). *Let $C_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ for $i \in [L]$ denote possibly different compression operators with compression coefficients γ_i. Let $\mathbf{x} = [\mathbf{x}_1\, \mathbf{x}_2\, \ldots \mathbf{x}_L]$, where $\mathbf{x}_i \in \mathbb{R}^{d_i}$ for all $i \in [L]$. Then $C(\mathbf{x}) := [C_1(\mathbf{x}_1)\, C_2(\mathbf{x}_2)\, \ldots C_L(\mathbf{x}_L)]$ is a compression operator with the compression coefficient being equal to $\gamma_{min} = \min_{i\in[L]} \gamma_i$.*

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

$$\begin{aligned}
\mathbb{E}_C\|\mathbf{x} - C(\mathbf{x})\|_2^2 &= \sum_{i=1}^{L} \mathbb{E}_{C_i}\|\mathbf{x}_i - C_i(\mathbf{x}_i)\|_2^2 \\
&\overset{\text{(a)}}{\leq} \sum_{i=1}^{L}(1-\gamma_i)\|\mathbf{x}_i\|_2^2 \\
&\leq (1-\gamma_{min})\|\mathbf{x}\|_2^2
\end{aligned}$$

Inequality (a) follows because each C_i is a compression operator with the compression coefficient γ_i. $\square$

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

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

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - \tfrac{Q_sComp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &= \|\mathbf{x}\|_2^2 - 2\mathbb{E}_C\left[\left\langle \mathbf{x}, \mathbb{E}_Q\left[\tfrac{Q_sComp_k(\mathbf{x})}{(1+\beta_{k,s})}\right]\right\rangle\right] + \mathbb{E}_{C,Q}\left[\tfrac{\|Q_sComp_k(\mathbf{x})\|_2^2}{(1+\beta_{k,s})^2}\right] \\
&\overset{\text{(a)}}{=} \|\mathbf{x}\|_2^2 - \tfrac{2}{(1+\beta_{k,s})}\mathbb{E}_C\left[\langle \mathbf{x}, Comp_k(\mathbf{x})\rangle\right] \\
&\quad + \tfrac{1}{(1+\beta_{k,s})^2}\mathbb{E}_{C,Q}\left[\|Q_sComp_k(\mathbf{x})\|_2^2\right]
\end{aligned}$$

In (a) we used $\mathbb{E}_Q[Q_sComp_k(\mathbf{x})] = Comp_k(\mathbf{x})$, which follows from (i) of Definition 1. Observe that, for $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$, we have $\langle \mathbf{x}, Comp_k(\mathbf{x})\rangle = \|Comp_k(\mathbf{x})\|_2^2$. Continuing from above, we get

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - \tfrac{Q_sComp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &= \|\mathbf{x}\|_2^2 - \tfrac{2}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\
&\quad + \tfrac{1}{(1+\beta_{k,s})^2}\mathbb{E}_{C,Q}\left[\|Q_sComp_k(\mathbf{x})\|_2^2\right] \qquad (13)
\end{aligned}$$

Observe that for any $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$, $Comp_k(\mathbf{x})$ is a length-d vector, but only (at most) k of its components are non-zero. This implies that, by treating $Comp_k(\mathbf{x})$ a length-k vector whose entries correspond to the k non-zero entries of $\mathbf{x}$, we can write $\mathbb{E}_Q[\|Q_sComp_k(\mathbf{x})\|_2^2] \leq (1+\beta_{k,s})\|Comp_k(\mathbf{x})\|_2^2$; see (ii) of Definition 1. Putting this back in (13), we get

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - \tfrac{Q_sComp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &\leq \|\mathbf{x}\|_2^2 - \tfrac{2}{1+\beta_{k,s}}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\
&\quad + \tfrac{1}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\
&= \|\mathbf{x}\|_2^2 - \tfrac{1}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \qquad (14)
\end{aligned}$$

Using $\mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] \geq \frac{k}{d}\|\mathbf{x}\|_2^2$ (see (16) in Lemma 5) in (14) gives

$$\begin{aligned}
\mathbb{E}_{C,Q}[\|\mathbf{x} - \tfrac{Q_sComp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &\leq \|\mathbf{x}\|_2^2 - \tfrac{(k/d)\|\mathbf{x}\|_2^2}{(1+\beta_{k,s})} \\
&= \left[1 - \tfrac{k}{d(1+\beta_{k,s})}\right]\|\mathbf{x}\|_2^2.
\end{aligned}$$

This completes the proof of Lemma 1.

$\square$

Lemma 5. *Let $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$. For any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}[\|Comp_k(\mathbf{x})\|_1^2] \geq \max\left\{\tfrac{k}{d}\|\mathbf{x}\|_2^2, \tfrac{k^2}{d^2}\|\mathbf{x}\|_1^2\right\} \tag{15}$$

$$\mathbb{E}[\|Comp_k(\mathbf{x})\|_2^2] \geq \tfrac{k}{d}\|\mathbf{x}\|_2^2. \tag{16}$$

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 $[d]$.

$$\begin{aligned}
\mathbb{E}[\|\text{Rand}_k(\mathbf{x})\|_m^2] &= \sum_{\omega\in\Omega_k} \tfrac{1}{|\Omega_k|}\left(\sum_{i=1}^d |x_i|^m \cdot \mathbb{1}\{i\in\omega\}\right)^{2/m} \\
&\stackrel{\text{(a)}}{\geq} \sum_{\omega\in\Omega_k} \tfrac{1}{|\Omega_k|}\sum_{i=1}^d |x_i|^2 \cdot \mathbb{1}\{i\in\omega\} \\
&= \sum_{i=1}^d x_i^2 \cdot \tfrac{1}{|\Omega_k|}\sum_{\omega\in\Omega_k} \mathbb{1}\{i\in\omega\} \\
&= \sum_{i=1}^d x_i^2 \cdot \tfrac{1}{|\Omega_k|}\binom{d-1}{k-1} \\
&= \tfrac{k}{d}\|\mathbf{x}\|_2^2
\end{aligned}$$

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

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

$\square$

Lemma 6. *For $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$, $\frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k}$, where $\|\cdot\|_m$ is the m-norm,*[12] *for any $m \in \mathbb{Z}_+$ is a compression operator with the compression coefficient γ_m being equal to*

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

Remark 1. *Note that this subsumes Lemma 2, which is for $m = 1$. 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 $\|Comp_k(\mathbf{x})\|_1 = \|Comp_k(\mathbf{x})\|_2$, 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.*

[12] The m-norm of a vector $\mathbf{u} \in \mathbb{R}^d$ is defined as $\|\mathbf{u}\|_m := \left(\sum_{i=1}^d |u_i|^m\right)^{\frac{1}{m}}$.

Proof of Lemma 6. Take an arbitrary $\mathbf{x} \in \mathbb{R}^d$.

$$
\begin{aligned}
&\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \left\langle \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k}, \mathbf{x} \right\rangle + \|\mathbf{x}\|_2^2 \right] \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \frac{\|Comp_k(\mathbf{x})\|_m \|Comp_k(\mathbf{x})\|_1}{k} + \|\mathbf{x}\|_2^2 \right] \\
&\leq \|\mathbf{x}\|_2^2 - \frac{\mathbb{E}_C \|Comp_k(\mathbf{x})\|_m^2}{k} \qquad (17)
\end{aligned}
$$

In (17) 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 (15)) in (17) gives

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

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 (17) with $q = m$ and $p = 2$ gives

$$
\begin{aligned}
&\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 \\
&\qquad \leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} \mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] \\
&\qquad \leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} (k/d) \|\mathbf{x}\|_2^2 \quad \text{(By Lemma 5)} \\
&\qquad = \left[1 - \frac{k^{\frac{2}{m}-1}}{d} \right] \|\mathbf{x}\|_2^2. \qquad (18)
\end{aligned}
$$

This completes the proof of Lemma 6. □

B Supplementary Material of Synchronous Qsparse-local-SGD from Section 3

B.1 Additional Theorem

Here we give a complementary result for Theorem 1, which was for a fixed learning rate. As noted earlier, the following theorem hold when Algorithm 1 is run with any compression operator (including our composed operators).

Theorem 5 (Convergence in the smooth (non-convex) case with decaying learning rate)**.** *Let $f^{(r)}(\mathbf{x})$ be L-smooth for every $r \in [R]$. Let $QComp_k : \mathbb{R}^d \to \mathbb{R}^d$ 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 $gap(\mathcal{I}_T) \leq H$, where $a > 1$ is such that, we have $a > \max\{\frac{4H}{\gamma}, 2\xi L, H\}$ and $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma - 4H}$. Then the following holds.*

$$
\mathbb{E}\|\nabla f(\mathbf{z}_T)\|^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2}{(a-1)P_T}\left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2} \right) + \left(\frac{8C}{\gamma^2} + 8 \right) \frac{\xi^3 L^2 G^2 H^2}{2(a-1)^2 P_T} \qquad (19)
$$

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.

B.2 Maintain Virtual Sequences

As outlined in Section 3, in order to prove our results, we define virtual sequences for every worker $r \in [R]$ and for all $t \geq 0$ as follows:

$$\widetilde{\mathbf{x}}_0^{(r)} := \widehat{\mathbf{x}}_0^{(r)} \qquad \text{and} \qquad \widetilde{\mathbf{x}}_{t+1}^{(r)} := \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \tag{20}$$

Define (i) $\mathbf{p}_t := \frac{1}{R}\sum_{r=1}^R \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$, $\quad \overline{\mathbf{p}}_t := \mathbb{E}_{i_t}[\mathbf{p}_t] = \frac{1}{R}\sum_{r=1}^R \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$;

and (ii) $\widetilde{\mathbf{x}}_{t+1} := \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t - \eta_t \mathbf{p}_t$, $\quad \widehat{\mathbf{x}}_t := \frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$

B.3 Bounding Error Compensation (Memory)

B.3.1 Difference of the true and the virtual parameter vectors is the average memory

Lemma 7 (Memory)**.** *The memory is maintained so as to capture the distance between the true sequence and virtual sequence.*

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R m_t^{(r)}. \tag{21}$$

Proof. Recall notation for an intermediate variable $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$ in Algorithm 1. Now consider $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t^{(r)}$. 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 $t'_r + 1 \leq t_r$

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^R \left(\widehat{\mathbf{x}}_{t_r+1}^{(r)} - \widetilde{\mathbf{x}}_{t_r+1}^{(r)}\right) \\
&= \frac{1}{R}\sum_{r=1}^R \left(\mathbf{x}_{t_r} - \frac{1}{R}\sum_{r=1}^R g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t'_r+1}^{(r)} - (\widehat{\mathbf{x}}_{t'_r+1}^{(r)} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)}))\right)
\end{aligned} \tag{22}$$

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

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^R \left(\mathbf{x}_{t_r} - \frac{1}{R}\sum_{r=1}^R g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t'_r+1}^{(r)} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)}))\right) \\
&= \mathbf{x}_{t'_r+1} - \frac{1}{R}\sum_{r=1}^R g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t'_r+1} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}})) \\
&= \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}) - \frac{1}{R}\sum_{r=1}^R g_{t_r}^{(r)}
\end{aligned} \tag{23}$$

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

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t_r} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}) - \frac{1}{R}\sum_{r=1}^R g_{t_r}^{(r)} \tag{24}$$

On rolling out the expression in (24) we get

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^R \left[\sum_{\substack{j:j+1\in\mathcal{I}_T \\ j\leq t_r}} \left(\mathbf{x}_j^{(r)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(r)} - g_j^{(r)}\right)\right] \\
&= \frac{1}{R}\sum_{r=1}^R m_{t_r+1}^{(r)} \\
&= \frac{1}{R}\sum_{r=1}^R m_t^{(r)}
\end{aligned} \tag{25}$$

Therefore $\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^{R} m_t^{(r)}$ is the average memory. $\square$

B.3.2 Memory contraction under the decaying learning rate

Lemma (Restating Lemma 4, Memory Contraction). *Let $\mathcal{I}_T \in [T]$ be a set of time instances in which the worker r updates and synchronizes with the master. For $a > \frac{4H}{\gamma}$, $\eta_t = \frac{\xi}{a+t}$, $gap(\mathcal{I}_T) \leq H$ and $t \in \mathbb{Z}^+$, there exists a $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma - 4H}$ such that*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta_t^2}{\gamma^2}CH^2G^2. \tag{26}$$

Proof. Fix an arbitrary worker $r \in [R]$. In order to prove the lemma, we need to show that $\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta_t^2}{\gamma^2}CH^2G^2$ 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 $\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2$. 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 $r \in [R]$.

$$\begin{aligned}
\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2 &= \mathbb{E}\|m_{t_{(i+1)}-1}^{(r)} + \mathbf{x}_{t_{(i+1)}-1} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)} - g_{t_{(i+1)}-1}^{(r)}\|^2 \\
&\overset{\text{(a)}}{\leq} (1-\gamma)\mathbb{E}\|m_{t_{(i+1)}-1}^{(r)} + \mathbf{x}_{t_{(i+1)}-1} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&\overset{\text{(b)}}{=} (1-\gamma)\,\mathbb{E}\|m_{t_{(i)}}^{(r)} + \mathbf{x}_{t_{(i)}} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&\overset{\text{(c)}}{=} (1-\gamma)\,\mathbb{E}\|m_{t_{(i)}}^{(r)} + \widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2
\end{aligned} \tag{27}$$

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 (27) gives (take any $p > 1$ in the following):

$$\begin{aligned}
\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2 &\leq (1-\gamma)\left[\left(1+\tfrac{(p-1)\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \left(1+\tfrac{p}{(p-1)\gamma}\right)\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2\right] \\
&\leq \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{(1-\gamma)(p\gamma+p)}{(p-1)\gamma}\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&= \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&= \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\mathbb{E}\|\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\leq \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\eta_{t_{(i)}}^2 H^2G^2
\end{aligned} \tag{28}$$

In the last inequality (28) we used $\mathbb{E}\|\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \leq \eta_{t_{(i)}}^2 H^2G^2$, which can be seen as follows:

$$\begin{aligned}
\mathbb{E}\|\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\eta_j \nabla^{(r)} f_{(i_j)}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 &= (t_{(i+1)} - t_{(i)})^2\mathbb{E}\|\tfrac{1}{(t_{(i+1)}-t_{(i)})}\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\overset{\text{(a)}}{\leq} (t_{(i+1)} - t_{(i)})\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\mathbb{E}\|\eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\overset{\text{(b)}}{\leq} (t_{(i+1)} - t_{(i)})\eta_{t_{(i)}}^2\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\mathbb{E}\|\nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2
\end{aligned}$$

$$\leq (t_{(i+1)} - t_{(i)})\eta_{t_{(i)}}^2 (t_{(i+1)} - t_{(i)})G^2$$
$$\overset{\text{(c)}}{\leq} \eta_{t_{(i)}}^2 H^2 G^2$$

Here (a) holds by Jensen's inequality, (b) holds since 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 (28) gives

$$\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2 \leq \left(1 - \tfrac{\gamma}{p}\right) \mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma} \tilde{\eta}_{t_{(i)}}^2 A. \tag{29}$$

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 $i = 1, 2, \ldots$, 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 $i = 1, 2, \ldots$. 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:

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

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 $\mathbb{E}\|m_{(i)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t_{(i)}}^2}{\gamma^2} A$ for some $i \in \mathbb{Z}^+$. We need to show that $\mathbb{E}\|m_{(i+1)}^{(r)}\|^2 \leq C \frac{\tilde{\eta}_{t_{(i+1)}}^2}{\gamma^2} A$. Using the inductive hypothesis in (29), we get

$$\begin{aligned}
\mathbb{E}\|m_{(i+1)}^{(r)}\|^2 &\leq \left(1 - \tfrac{\gamma}{p}\right) C \frac{\tilde{\eta}_{t_{(i)}}^2}{\gamma^2} A + \tfrac{p(1-\gamma^2)}{(p-1)\gamma} \tilde{\eta}_{t_{(i)}}^2 A \\
&= C \frac{\tilde{\eta}_{t_{(i)}}^2}{\gamma^2} A \left(1 - \tfrac{\gamma}{p} + \tfrac{p(1-\gamma^2)}{p-1} \tfrac{\gamma}{C}\right) \\
&= C \frac{\tilde{\eta}_{t_{(i)}}^2}{\gamma^2} A \left(1 - \tfrac{\gamma}{p}\left(1 - \tfrac{p^2(1-\gamma^2)}{(p-1)C}\right)\right)
\end{aligned} \tag{30}$$

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 $\beta \geq 2H$.

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

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

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 (30) and Claim 1 implies that if $C \geq \frac{4\gamma a(1-\gamma^2)}{(a\gamma-4H)}$, where $a > 4H/\gamma$, then $\mathbb{E}\|m_{(i+1)}^{(r)}\|^2 \leq C\frac{\tilde{\eta}_{t_{(i+1)}}^2}{\gamma^2}A$ 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. □

B.3.3 Bounded memory under the fixed learning rate

Lemma (Restating Lemma 3, Bounded Memory)**.** *Let $\mathcal{I}_T^{(r)} \in [T]$ be a set of time instances in which the worker r updates and synchronizes with the master. For $\eta_t = \eta$, $gap(\mathcal{I}_T) \leq H$ and $t \in \mathbb{Z}^+$ we have*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta^2(1-\gamma^2)}{\gamma^2}H^2G^2 \tag{31}$$

Proof. Observe that (28) holds irrespective of the learning rate schedule. In particular, using a fixed learning rate $\eta_t = \eta$ for every t gives

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

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

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

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

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

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

B.4 Sequence Deviation

B.4.1 Contracting deviation of local true sequences from the global true sequence under decaying learning rate

Lemma 8 (Contracting Deviation of Local Sequences)**.** *Similar to Lemma 3.3 in [31] we bound the deviation of the local sequences.*

$$\tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2 \tag{32}$$

Proof. 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 $\mathbf{u}_1,\ldots,\mathbf{u}_R$, if we let $\bar{\mathbf{u}} = \frac{1}{R}\sum_{i=1}^{r}\mathbf{u}_i$, then $\sum_{i=1}^{n}\|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 \leq \sum_{i=1}^{R}\|\mathbf{u}_i\|^2$. We use this in the first inequality below.

$$\begin{aligned}
\tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &= \tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)} - (\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_{t_r}^{(r)})\|^2 \\
&\leq \tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \\
&\leq \eta_{t_r}^2 G^2 H^2 \\
&\leq 4\eta_t^2 G^2 H^2
\end{aligned} \tag{33}$$

The last inequality (33) uses $\eta_{t_r} \leq 2\eta_{t_r+H} \leq 2\eta_t$ and $t - t_r \leq H$. □

B.4.2 Bounded deviation of local true sequences from the global true sequence under fixed learning rate

Lemma 9 (Bounded Deviation of Local Sequences)**.** *With $\eta_t = \eta$ this follows from the analysis of Lemma 8*

$$\tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \eta^2 G^2 H^2 \tag{34}$$

Proof. Similar to analysis in (33) we can show that $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \eta^2 G^2 H^2$. □

B.5 Smooth and Non-Convex Objective

B.5.1 Proof of Theorem 1 – Fixed Learning Rate

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 5, we start off with the decaying learning rate η_t until (38) and then switch to the fixed learning rate η. Note that the proof remains the same until (38) irrespective of the learning rate schedule; in particular, we can take $\eta_t = \eta$ and the same proof holds until (38).

By the definition of L-smoothness, we have

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

Define i_t as the set of random sampling of the mini-batches at each worker $\{i_t^{(1)}, i_t^{(2)}, \ldots, i_t^{(R)}\}$. 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

$$
\begin{aligned}
\mathbb{E}_{i_t}[f(\widetilde{\mathbf{x}}_{t+1})] - f(\widetilde{\mathbf{x}}_t) &\le -\tfrac{\eta_t}{2}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 + \|\tfrac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 - \|\nabla f(\widetilde{\mathbf{x}}_t) - \tfrac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 \right) \\
&\qquad + \eta_t^2 L \|\tfrac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 + \tfrac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 \\
&\le -\tfrac{\eta_t}{2R}\sum_{r=1}^{R}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 - L^2\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \right) + \tfrac{2\eta_t^2 L - \eta_t}{2}\|\tfrac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 \\
&\qquad + \tfrac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 \\
&= -\tfrac{\eta_t}{2R}\sum_{r=1}^{R}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 + L^2\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \right) + \tfrac{2\eta_t^2 L - \eta_t}{2R}\sum_{r=1}^{R}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \\
&\qquad + \tfrac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + \tfrac{\eta_t L^2}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2. \qquad (35)
\end{aligned}
$$

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

$$
\begin{aligned}
\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\le 2\|\nabla f(\widehat{\mathbf{x}}_t^{(r)}) - \nabla f(\widetilde{\mathbf{x}}_t)\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 \\
&\le 2L^2\|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2, \qquad (36)
\end{aligned}
$$

where the 2nd inequality follows from the smoothness (L-Lipschitz gradient) assumption. Using this and that $\eta_t \le \frac{1}{2L}$ in (35) and rearranging terms give

$$
\tfrac{\eta_t}{4R}\sum_{r=1}^{R}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \le f(\widetilde{\mathbf{x}}_t) - \mathbb{E}_{(i_t)}[f(\widetilde{\mathbf{x}}_{t+1})] + \tfrac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + \tfrac{\eta_t L^2}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \qquad (37)
$$

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

$$
\begin{aligned}
\tfrac{\eta_t}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\le \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \tfrac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 2\eta_t L^2\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 \\
&\qquad + 2\eta_t L^2 \tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \qquad (38)
\end{aligned}
$$

Observe that (38) holds irrespective of the learning rate schedule. In particular, if we take a fixed learning rate $\eta_t = \eta \le \frac{1}{2L}$ in (38), we get

$$
\begin{aligned}
\tfrac{\eta}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\le \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \tfrac{\eta^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 2\eta L^2\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 \\
&\qquad + 2\eta L^2 \tfrac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \qquad (39)
\end{aligned}
$$

Lemma 7 and Lemma 3 together imply $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \le \frac{4\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2$. We also have from Lemma 9 that $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \le \eta^2 G^2 H^2$. Substituting these in (39) gives

$$
\begin{aligned}
\tfrac{\eta}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\le \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \tfrac{\eta^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 8\tfrac{\eta^3(1-\gamma^2)}{\gamma^2}L^2G^2H^2 \\
&\qquad + 2\eta^3 L^2 G^2 H^2 \qquad (40)
\end{aligned}
$$

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

$$\frac{1}{4RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \frac{\mathbb{E}[f(\widetilde{\mathbf{x}}_0)]-f^*}{\eta T} + \frac{\eta L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 8\frac{\eta^2(1-\gamma^2)}{\gamma^2}L^2G^2H^2 + 2\eta^2L^2G^2H^2 \tag{41}$$

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

$$\frac{1}{RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \left(\frac{\mathbb{E}[f(\mathbf{x}_0)]-f^*}{\widehat{C}} + \frac{\widehat{C}L}{bR^2}\sum_{r=1}^{R}\sigma_r^2\right)\frac{4}{\sqrt{T}} + 8\left(4\frac{(1-\gamma^2)}{\gamma^2}+1\right)\frac{\widehat{C}^2L^2G^2H^2}{T}. \tag{42}$$

Sample a parameter $\mathbf{z}_T$ from $\left\{\widehat{\mathbf{x}}_t^{(r)}\right\}$ 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 (42) gives

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

This completes the proof of Theorem 1. □

B.5.2 Proof of Theorem 5 – Decaying Learning Rate

Proof. Observe that we can use the proof of Theorem 1 exactly until (38), for $\eta_t \leq \frac{1}{2L}$ (which follows from our assumption that $a \geq 2\xi L$), which gives

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

We have from Lemma 8 that $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2G^2H^2$. Lemma 7 and Lemma 4 together imply that $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \frac{1}{R}\sum_{r=1}^{R}\|m_t^{(r)}\|^2 \leq C\frac{4\eta_t^2}{\gamma^2}G^2H^2$. Using these bounds in (43) gives

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

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

$$\sum_{t=0}^{T-1}\frac{\eta_t}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\mathbf{x}_0)] - f^* + \frac{L\sum_{r=1}^{R}\sigma_r^2}{bR^2}\sum_{t=0}^{T-1}\eta_t^2 + \left(\frac{8C}{\gamma^2}+8\right)L^2G^2H^2\sum_{t=0}^{T-1}\eta_t^3. \tag{44}$$

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 (44) yields

$$\frac{1}{P_T}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\delta_t\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0)-f^*}{P_T} + \frac{L\xi^2}{bR^2(a-1)}\frac{\sum_{r=1}^{R}\sigma^2}{P_T} + \left(\frac{8C}{\gamma^2}+8\right)L^2G^2H^2\frac{\xi^3}{2P_T(a-1)^2} \tag{45}$$

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

$$\min_{t\in\{0,\dots,T-1\},\, r\in[R]}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \xrightarrow{T\to\infty} 0. \tag{46}$$

Sample a parameter $\mathbf{z}_T$ from $\left\{\widehat{\mathbf{x}}_t^{(r)}\right\}$ for $r = 1, \ldots, R$ and $t = 0, 1, \ldots, T-1$ 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 (45)

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

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

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

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

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

This completes the proof of Theorem 5. □

B.6 Smooth and Strongly Convex Objective: Proof of Theorem 2 – Decaying Learning Rate

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 $\widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$ for each worker $r \in [R]$ and defining $\mathbf{p}_t := \frac{1}{R}\sum_{r=1}^{R}\nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$, we get

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{\mathbf{x}}_t - \eta_t \mathbf{p}_t. \tag{47}$$

Define i_t as the set of random sampling of the mini-batches at each worker $\{i_t^{(1)}, i_t^{(2)}, \ldots, i_t^{(R)}\}$ and let $\overline{\mathbf{p}}_t = \mathbb{E}_{i_t}[\mathbf{p}_t]$. From (47) we can get

$$\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 - 2\eta_t\left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t, \mathbf{p}_t - \overline{\mathbf{p}}_t\right\rangle \tag{48}$$

Taking the expectation w.r.t. the sampling i_t at time t (conditioning on the past) and noting that last term in (48) becomes zero gives:

$$\mathbb{E}_{i_t}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\mathbb{E}_{i_t}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 \tag{49}$$

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

$$\mathbb{E}_{i_t}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\frac{\sum_{r=1}^{R}\sigma_r^2}{bR^2}. \tag{50}$$

Now we bound the first term on the RHS.

Lemma 10. *If $\eta_t \leq \frac{1}{4L}$, then we have*

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 \leq \left(1 - \frac{\mu\eta_t}{2}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t\mu}{2L}(f(\widehat{\mathbf{x}}_t) - f^*)$$
$$+\,\eta_t\left(\frac{3\mu}{2}+3L\right)\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \tag{51}$$

Proof.

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 + \eta_t^2\|\overline{\mathbf{p}}_t\|^2 - 2\eta_t\left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t\right\rangle \tag{52}$$

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

$$\|\overline{\mathbf{p}}_t\|^2 = \|\tfrac{1}{R}\sum_{r=1}^{R}\left(\nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - \nabla f^{(r)}(\widetilde{\mathbf{x}}_t)\right) + \nabla f(\widetilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*)\|^2$$
$$\leq \tfrac{1}{R}\sum_{r=1}^{R} 2\|\nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - \nabla f^{(r)}(\widetilde{\mathbf{x}}_t)\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t) - \nabla f\left(\mathbf{x}^*\right)\|^2$$
$$\leq \tfrac{2L^2}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\| + 2\|\nabla f(\widetilde{\mathbf{x}}_t) - \nabla f\left(\mathbf{x}^*\right)\|^2 \tag{53}$$

By the definition of smoothness, we have $\|\nabla f\left(\widetilde{\mathbf{x}}_t\right) - \nabla f\left(\mathbf{x}^*\right)\|^2 \leq 2L\left(f\left(\widetilde{\mathbf{x}}_t\right) - f(\mathbf{x}^*)\right)$, where $\nabla f(\mathbf{x}^*) = 0$. Substituting this in (53) gives

$$\eta_t^2\|\overline{\mathbf{p}}_t\|^2 \leq \tfrac{2\eta_t^2 L^2}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\| + 4\eta_t^2 L\left(f\left(\widetilde{\mathbf{x}}_t\right) - f(\mathbf{x}^*)\right) \tag{54}$$

Now we bound the last term of (52). By definition, we have

$$-2\eta_t\left\langle\widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t\right\rangle = -2\tfrac{\eta_t}{R}\sum_{r=1}^{R}\left\langle\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)\right\rangle - 2\tfrac{\eta_t}{R}\sum_{r=1}^{R}\left\langle\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)\right\rangle \tag{55}$$

For the first term on the RHS of (55), we can use strong convexity

$$-2\left\langle\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)\right\rangle \leq -2\left(f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - f^{(r)}(\mathbf{x}^*)\right) - \mu\|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \tag{56}$$

For the second term on the RHS of (55), we can use the following by smoothness.

$$-2\left\langle\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)\right\rangle \leq L\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 + 2\left(f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - f^{(r)}\left(\widetilde{\mathbf{x}}_t\right)\right) \tag{57}$$

Using (56)-(57) in (55) we get

$$-2\eta_t\left\langle\widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t\right\rangle \leq -\tfrac{2\eta_t}{R}\sum_{r=1}^{R}\left(f^{(r)}\left(\widetilde{\mathbf{x}}_t\right) - f^{(r)}(\mathbf{x}^*)\right) - \tfrac{\eta_t\mu}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 + \tfrac{L\eta_t}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2$$
$$= -2\eta_t\left(f\left(\widetilde{\mathbf{x}}_t\right) - f(\mathbf{x}^*)\right) - \tfrac{\eta_t\mu}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 + L\tfrac{\eta_t}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \tag{58}$$

Adding (54) and (58) and using $a \geq 32L/\mu$ which implies $\eta_t \leq {}^1\!/_{4L}$ yields

$$\eta_t^2\|\overline{\mathbf{p}}_t\|^2 - 2\eta_t\left\langle\widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t\right\rangle \leq -2\eta_t(1 - 2\eta_t L)\left(f\left(\widetilde{\mathbf{x}}_t\right) - f^*\right) - \tfrac{\eta_t\mu}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2$$
$$+ \tfrac{L\eta_t + 2\eta_t^2 L^2}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2$$
$$\leq -\eta_t\left(f\left(\widetilde{\mathbf{x}}_t\right) - f^*\right) - \eta_t\mu\|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2$$
$$+ \tfrac{3L\eta_t}{R}\sum_{r=1}^{R}\left(\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 + \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2\right) \tag{59}$$

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

$$-\|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \leq \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 - \tfrac{1}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \tag{60}$$

Using (60) in (59) and then substituting (59) in (52) gives

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 \leq \left(1 - \tfrac{\mu\eta_t}{2}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \eta_t\left(f\left(\widetilde{\mathbf{x}}_t\right) - f^*\right)$$
$$+ \eta_t\left(\mu + 3L\right)\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \tfrac{3L\eta_t}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \tag{61}$$

Using strong convexity of f we have

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 \leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\eta_t \mu}{2} \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 + \eta_t \left(\mu + 3L\right) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \tfrac{3L\eta_t}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \tag{62}$$

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

$$\begin{aligned}
\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 &\leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\eta_t \mu}{4} \|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \eta_t \left(\tfrac{3\mu}{2} + 3L\right) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \tfrac{3L\eta_t}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \\
&\leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\eta_t \mu}{2L} \left(f(\widehat{\mathbf{x}}_t) - f^*\right) \quad \text{(Using smoothness of } f(\mathbf{x})) \\
&\quad + \eta_t \left(\tfrac{3\mu}{2} + 3L\right) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \tfrac{3L\eta_t}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2
\end{aligned} \tag{63}$$

This completes the proof of Lemma 10. □

Using (63) in (50) and then taking the expectation over the entire process gives

$$\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\eta_t\mu}{2L}\left(\mathbb{E}[f(\widehat{\mathbf{x}}_t)] - f^*\right) + \eta_t \left(\tfrac{3\mu}{2} + 3L\right) \mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R} \sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 + \eta_t^2 \tfrac{\sum_{r=1}^R \sigma_r^2}{bR^2} \tag{64}$$

From Lemma 8, we have $\frac{1}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2$. Lemma 7 and Lemma 4 together imply that $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq 4C\frac{\eta_t^2}{\gamma^2}H^2G^2$. Substituting these back in (64) and letting $e_t = \mathbb{E}[f(\widehat{\mathbf{x}}_t) - f^*]$ gives

$$\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\mu\eta_t}{2L} e_t + \eta_t \left(\tfrac{3\mu}{2} + 3L\right) C \tfrac{4\eta_t^2}{\gamma^2} G^2 H^2 + (3L\eta_t) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \tfrac{\sum_{r=1}^R \sigma_r^2}{bR^2} \tag{65}$$

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

$$\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \left(1 - \tfrac{\mu\eta_t}{2}\right) \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \tfrac{\mu\eta_t}{2L} e_t + \eta_t \left(\tfrac{3\mu}{2} + 3L\right) C \tfrac{4\eta_t^2}{\gamma^2} G^2 H^2 + (3\eta_t L) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \tfrac{\sum_{r=1}^R \sigma_r^2}{bR^2} \tag{66}$$

Employing a slightly modified Lemma 3.3 from [30] with $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$, $B = 4\left(\left(\frac{3\mu}{2} + 3L\right)\frac{CG^2H^2}{\gamma^2} + 3L^2G^2H^2\right)$, and $a_t = \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2$, we have

$$a_{t+1} \leq \left(1 - \tfrac{\mu\eta_t}{2}\right) a_t - \tfrac{\mu\eta_t}{2L} e_t + \eta_t^2 A + \eta_t^3 B \tag{67}$$

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

$$\tfrac{\mu}{2LS_T} \sum_{t=0}^{T-1} w_t e_t \leq \tfrac{\mu a^3}{8S_T} a_0 + \tfrac{4T(T+2a)}{\mu S_T} A + \tfrac{64T}{\mu^2 S_T} B \tag{68}$$

From convexity we can finally write

$$\mathbb{E} f\left(\overline{\mathbf{x}}_T\right) - f^* \leq \tfrac{La^3}{4S_T} a_0 + \tfrac{8LT(T+2a)}{\mu^2 S_T} A + \tfrac{128LT}{\mu^3 S_T} B \tag{69}$$

Where $\overline{\mathbf{x}}_T := \frac{1}{S_T}\sum_{t=0}^{T-1}\left[w_t\left(\frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}\right)\right] = \frac{1}{S_T}\sum_{t=0}^{T-1} w_t \widehat{\mathbf{x}}_t$ □

Algorithm 2 Qsparse-local-SGD with asynchronous updates

Initialize $\mathbf{x}_0 = \bar{\bar{\mathbf{x}}}_0 = \mathbf{x}_0^{(r)} = \widehat{\mathbf{x}}_0^{(r)} = m_0^{(r)} = \mathbf{0},\ \forall r \in [R]$. Suppose η_t follows a certain learning rate schedule.
for $t = 0$ **to** $T-1$ **do**
  On Workers:
  for $r = 1$ **to** R **do**
    $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} \leftarrow \widehat{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$; $i_t^{(r)}$ is a mini-batch of size b uniformly in $\mathcal{D}_r$
    if $t+1 \notin \mathcal{I}_T^{(r)}$ **then**
      $\mathbf{x}_{t+1}^{(r)} \leftarrow \mathbf{x}_t^{(r)}$, $m_{t+1}^{(r)} \leftarrow m_t^{(r)}$ and $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$
    else
      $g_t^{(r)} \leftarrow QComp_k\left(m_t^{(r)} + \mathbf{x}_t^{(r)} - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}\right)$ and send $g_t^{(r)}$ to the master
      $m_{t+1}^{(r)} \leftarrow m_t^{(r)} + \mathbf{x}_t^{(r)} - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)} - g_t^{(r)}$
      Receive $\bar{\bar{\mathbf{x}}}_{t+1}$ from the master and set $\mathbf{x}_{t+1}^{(r)} \leftarrow \bar{\bar{\mathbf{x}}}_{t+1}$ and $\widehat{\mathbf{x}}_{t+1}^{(r)} \leftarrow \bar{\bar{\mathbf{x}}}_{t+1}$
    end if
  end for
  At Master:
  if $t+1 \notin \mathcal{I}_T^{(r)}$ for all $r \in [R]$ **then**
    $\bar{\bar{\mathbf{x}}}_{t+1} \leftarrow \bar{\bar{\mathbf{x}}}_t$
  else
    Let $\mathcal{S} \subseteq [R]$ be the set of all workers r such that master receives $g_t^{(r)}$ from r.
    Compute $\bar{\bar{\mathbf{x}}}_{t+1} \leftarrow \bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r\in\mathcal{S}} g_t^{(r)}$ and broadcast $\bar{\bar{\mathbf{x}}}_{t+1}$ to all the workers in $\mathcal{S}$.
  end if
end for

C Supplementary Material of Asynchronous Qsparse-local-SGD from Section 4

Our algorithm for Asynchronous Qsparce-local-SGD is given below.

Below we give more precise statements of Theorem 3 and Theorem 4 from Section 4 (see Theorem 6 and Theorem 7, respectively), along with an additional theorem (see Theorem 8), which is a complementary result for Theorem 3, which was for a fixed learning rate. As noted earlier, the following theorem hold when Algorithm 2 is run with any compression operator (including our composed operators).

Theorem 6 (Convergence in the smooth and non-convex case with fixed learning rate)**.** *Under the same conditions as in Theorem 1 with $gap(\mathcal{I}_T^{(r)}) \leq H$ and $C_1 = (\frac{8}{\gamma^2} - 6)(4 - 2\gamma)$, if $\{\widehat{x}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to Algorithm 2, the following holds.*

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

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

Corollary. *Under the same conditions as in Theorem 1 with $gap(\mathcal{I}_T^{(r)}) \leq H$, if $\{\widehat{\mathbf{x}}_t^{(r)}\}_{t=0}^{T-1}$ is generated according to Algorithm 2, the following holds, where $\mathbb{E}[f(\mathbf{x}_0)] - f^* \leq J^2$, $\sigma_{max} = \max_{r\in[R]}\sigma_r$, and $\widehat{C}^2 = {bR(\mathbb{E}[f(\mathbf{x}_0)]-f^*)}/{\sigma_{max}^2}$.*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|^2 \leq \mathcal{O}\left(\frac{J\sigma_{max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2bRG^2}{\sigma_{max}^2\gamma^2T}(H^2 + H^4)\right), \tag{70}$$

where $\mathbf{z}_T$ is a random variable which samples a previous parameter $\widehat{\mathbf{x}}_t^{(r)}$ with probability $1/RT$. 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 $H = \mathcal{O}\left(\sqrt{\gamma}T^{1/8}/(bR)^{3/8}\right)$.

Theorem 6 provides non asymptotic guarantees where we also observe that the compression comes for "free". The corresponding asymptotic result has been omitted to Appendix C.

Theorem 7 (Convergence in the smooth and strongly convex case with decaying learning rate)**.** *Under the same conditions as in Theorem 2 with gap*$(\mathcal{I}_T^{(r)}) \leq H$*, if* $\{\widehat{x}_t^{(r)}\}_{t=0}^{T-1}$ *is generated according to Algorithm 2, the following holds.*

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \frac{La^3}{4S_T}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{8LT(T+2a)}{\mu^2 S_T}A + \frac{128LT}{\mu^3 S_T}D \tag{71}$$

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) $A = \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}$, $D = \left(\frac{3\mu}{2} + 3L\right)\left(\frac{12CG^2H^2}{\gamma^2} + C_1\eta_t^2 H^4 G^2\right) + 24(1+C_2H^2)LG^2H^2$*; and* (iii) $\overline{\mathbf{x}}_T$*,* S_T *are as defined in Theorem 2.*

Corollary. *Under the same conditions as in Theorem 2 with 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:*

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \mathcal{O}\left(\frac{G^2H^3}{\mu^2\gamma^3T^3}\right) + \mathcal{O}\left(\frac{\sigma_{max}^2}{\mu^2 bRT} + \frac{H\sigma_{max}^2}{\mu^2 bR\gamma T^2}\right) + \mathcal{O}\left(\frac{G^2}{\mu^3\gamma^2T^2}(H^2+H^4)\right), \tag{72}$$

where $\overline{\mathbf{x}}_T$*,* S_T *are as defined in Theorem 2. 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* $H = \mathcal{O}\left(\sqrt{\gamma}(T/(bR))^{1/4}\right)$*.*

C.1 Additional Theorem

Theorem 8 (Convergence in the smooth and non-convex case with decaying learning rate)**.** *Let* $f^{(r)}(\mathbf{x})$ *be L-smooth for every* $r \in [R]$*. Let* $QComp_k : \mathbb{R}^d \to \mathbb{R}^d$ *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* $\eta_t = \frac{\xi}{(a+t)}$*, gap*$(\mathcal{I}_T^r) \leq H$ *for* $r \in [R]$*, where* $a > 1$ *is such that, we have* $a > \max\{\frac{4H}{\gamma}, 2\xi L, H\}$*,*$C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$*. Then for* $C' = (4-2\gamma)(1+\frac{C}{\gamma^2})$ *the following holds.*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0) - f^*}{P_T} + \frac{L\xi^2}{(a-1)P_T}\left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2}\right) + \left(16 + \frac{24C}{\gamma^2} + 200C'H^2\right)\frac{\xi^3L^2G^2H^2}{2(a-1)^2P_T} \tag{73}$$

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*.*

C.2 Maintain Virtual Sequences

As noted earlier in Section 3 and also in Appendix B, 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:

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

Define

1. $\widetilde{\mathbf{x}}_{t+1} := \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t+1}^{(r)} = \widetilde{\mathbf{x}}_t - \frac{\eta_t}{R}\sum_{r=1}^R \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$
2. $\mathbf{p}_t := \frac{1}{R}\sum_{r=1}^R \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$
3. $\overline{\mathbf{p}}_t := \mathbb{E}_{(i_t)}[\mathbf{p}_t] = \frac{1}{R}\sum_{r=1}^R \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right)$
4. $\widehat{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_t^{(r)}$
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.3 Contracting local sequence deviation under decaying learning rate

Lemma 11 (Contracting Local Sequence Deviation)**.** *For* $\widehat{\mathbf{x}}_t, \widehat{\mathbf{x}}_t^{(r)}$ *generated according to Algorithm 2 and gap*$(\mathcal{I}_T^{(r)}) \leq H$ *the following holds*

$$\frac{1}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 8(1+C''H^2)\eta_t^2G^2H^2 \tag{74}$$

Here $C'' = 8B(1+\frac{C}{\gamma^2})$ *where* B *is from* $\mathbb{E}_{Q,C}\|QComp_k(\mathbf{x})\|^2 \leq B\|\mathbf{x}\|^2$ *and* $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$

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 $\mathbf{u}_1, \ldots, \mathbf{u}_R$, if we let $\bar{\mathbf{u}} = \frac{1}{R}\sum_{i=1}^{r} \mathbf{u}_i$, then $\sum_{i=1}^{n} \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 \le \sum_{i=1}^{R} \|\mathbf{u}_i\|^2$. We use this in the first inequality below.

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &= \frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'} - (\widehat{\mathbf{x}}_t - \bar{\bar{\mathbf{x}}}_{t_0'})\|^2 \\
&\le \frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \\
&\le \frac{2}{R}\sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 + \frac{2}{R}\sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \qquad (75)
\end{aligned}$$

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

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

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

$$\bar{\bar{\mathbf{x}}}_{t_r}^{(r)} = \bar{\bar{\mathbf{x}}}_{t_0'} - \frac{1}{R}\sum_{s=1}^{R}\sum_{j=t_0'}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(s)}\} g_j^{(s)}. \qquad (77)$$

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

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \le \frac{(t_r - t_0')}{R}\sum_{s=1}^{R}\sum_{j=t_0'}^{t_r} \mathbb{E}\|g_j^{(s)}\|^2 \qquad (78)$$

Now we bound $\mathbb{E}\|g_j^{(s)}\|^2$ for any $j \in \{t_0', \ldots, t_r\}$ and $s \in [R]$: Since $\mathbb{E}\|QC(\mathbf{u})\|^2 \le B\|\mathbf{u}\|^2$ holds for every $\mathbf{u}$, where $B = (4 - 2\gamma)$,[13] we have for any $s \in [R]$ that

$$\begin{aligned}
\mathbb{E}\|g_j^{(s)}\|^2 &\le B\mathbb{E}\|m_j^{(s)} + \mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \qquad (79) \\
&\le 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \qquad (80)
\end{aligned}$$

Observe that the proof of Lemma 4 does not depend on the synchrony of the network; it only uses the fact that $gap(\mathcal{I}_T^{(s)}) \le H$ for any worker $s \in [R]$. Therefore, we can directly use Lemma 4 to bound the first term in (76) as $\mathbb{E}\|m_j^{(s)}\|^2 \le 4C\frac{\eta_j^2}{\gamma^2}H^2G^2$. In order to bound the second term of (76), note that $\mathbf{x}_j^{(s)} = \widehat{\mathbf{x}}_{t_s}^{(s)}$, which implies that $\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 = \|\sum_{l=t_s}^{j} \eta_l \nabla f_{i_l^{(s)}}\left(\widehat{\mathbf{x}}_l^{(s)}\right)\|^2$. Taking expectation yields $\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \le 4\eta_{t_s}^2 H^2G^2 \le 4\eta_{t_0'}^2 H^2 G^2$, where in the second inequality we used $t_0' \le t_s$, which implies $\eta_{t_s} \le \eta_{t_0'}$. Using these in (80) gives

$$\mathbb{E}\|g_j^{(s)}\|^2 \le 8B\left(1 + \frac{C}{\gamma^2}\right)\eta_{t_0'}^2 H^2 G^2. \qquad (81)$$

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

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 $\eta_{n'_0} \leq 2\eta_t$ in (81)) in (78) gives

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

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 11. □

C.4 Bounded local sequence deviation under fixed learning rate

Lemma 12 (Bounded Local Sequence Deviation). *For $\widehat{\mathbf{x}}_t, \widehat{\mathbf{x}}_t^{(r)}$ generated according to Algorithm 2 with $\eta_t = \eta$ the following holds*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq (2 + H^2 C')\eta^2 G^2 H^2 \tag{83}$$

Here $C' = (\frac{16}{\gamma^2} - 12)B$ where B is from $\mathbb{E}_{Q,C}\|QComp_k(\mathbf{x})\|^2 \leq B\|\mathbf{x}\|^2$.

Proof. From (79) and (80) and using the fact that for a given QC operator, we show that $\mathbb{E}\|QC(\mathbf{u})\|^2 \leq B\|\mathbf{u}\|^2$ holds for every $\mathbf{u}$

$$\begin{aligned}\mathbb{E}\|g_j^{(s)}\|^2 &\leq 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\eta^2 H^2 G^2 \\ &\leq 8B\frac{(1-\gamma^2)\eta^2}{\gamma^2}H^2G^2 + 2\eta^2 BH^2G^2 \\ &= 2B\left(\frac{4}{\gamma^2} - 3\right)\eta^2 H^2 G^2\end{aligned} \tag{84}$$

For a fixed learning rate η, using (84) and following similar analysis as in (76) we can bound the first term in (75) as follows

$$\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \leq \eta^2 H^2 G^2 \tag{85}$$

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

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t'_0}\|^2 \leq 2B\left(\frac{4}{\gamma^2} - 3\right)\eta^2 H^4 G^2 \tag{86}$$

Using (85) and (86) in (75) we can show that

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \left[2 + 4BH^2\left(\frac{4}{\gamma^2} - 3\right)\right]\eta^2 H^2 G^2 \tag{87}$$

□

C.5 Contracting distance between virtual and true sequence under decaying learning rate

Lemma 13 (Contracting distance between Virtual and True Sequence). *Let $\mathcal{I}_T^{(r)} \in [T]$ be a set of time instances in which the worker r updates and synchronizes with the master. For $a > \frac{4H}{\gamma}$, $\eta_t = \frac{\xi}{a+t}$, $gap(\mathcal{I}_T^{(r)} \leq H)$ and $t \in \mathbb{Z}^+$, there exists a $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma - 4H}$ such that*

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq C'\eta_t^2 H^4 G^2 + 12C\frac{\eta_t^2}{\gamma^2}G^2 H^2 \tag{88}$$

Here $C' = 192B\left(1 + \frac{C}{\gamma^2}\right)$ where B is from $\mathbb{E}_{Q,C}\|QComp_k(\mathbf{x})\|^2 \leq B\|\mathbf{x}\|^2$.

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 7 that $\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R m_t^{(r)}$. 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

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \left[\frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\right] + \left[\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\right] + \left[\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)}\right] \tag{89}$$

Applying Jensen's inequality and taking expectation gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \left[\frac{3}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2\right] + \left[3\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2\right] + \left[3\mathbb{E}\|\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)}\|^2\right] \tag{90}$$

We bound each of the three terms of (90) separately. We have upper-bounded the first term earlier in (82), which is

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \leq 32B\left(1 + \frac{C}{\gamma^2}\right)\eta_t^2 H^4 G^2. \tag{91}$$

To bound the second term of (90), note that

$$\bar{\bar{\mathbf{x}}}_t = \bar{\bar{\mathbf{x}}}_0 - \frac{1}{R}\sum_{r=1}^R \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \tag{92}$$

$$= \bar{\bar{\mathbf{x}}}_{t_0'} - \frac{1}{R}\sum_{r=1}^R \sum_{j=t_0'}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \tag{93}$$

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 (93) as

$$\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2 \leq \frac{(t_r - t_0')}{R}\sum_{r=1}^R \sum_{j=t_0'}^{t_r} \mathbb{E}\|g_j^{(r)}\|^2$$

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

$$\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2 \leq 32B\left(1 + \frac{C}{\gamma^2}\right)\eta_t^2 H^4 G^2. \tag{94}$$

To bound the last term of (90), note that

$$\widetilde{\mathbf{x}}_{t_r}^{(r)} = \bar{\bar{\mathbf{x}}}_0 - \sum_{j=0}^{t_r-1} \eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right) \tag{95}$$

From (92) and (95), we can write

$$\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)} = \frac{1}{R}\sum_{r=1}^R \left[\sum_{j=0}^{t_r-1} \eta_j \nabla^{(r)} f_{(i_j)}\left(\widehat{\mathbf{x}}_j^{(r)}\right) - \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)}\right] \tag{96}$$

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

$$\mathbf{x}_{t_r^{(2)}-1}^{(r)} - \widehat{\mathbf{x}}_{t_r^{(2)}-\frac{1}{2}}^{(r)} = \sum_{j=t_r^{(1)}}^{t_r^{(2)}-1} \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right). \tag{97}$$

Using (97) for every consecutive synchronization steps, we can equivalently write (96) as

$$\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^{R}\widetilde{\mathbf{x}}_{t_r}^{(r)} = \frac{1}{R}\sum_{r=1}^{R}\left[\sum_{\substack{j:j+1\in\mathcal{I}_T^{(r)}\\ j\le t_r-1}}\left(\mathbf{x}_j^{(r)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(r)} - g_j^{(r)}\right)\right]$$

$$= \frac{1}{R}\sum_{r=1}^{R} m_{t_r}^{(r)}$$

$$= \frac{1}{R}\sum_{r=1}^{R} m_t^{(r)} \tag{98}$$

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 11, we can directly apply Lemma 4 to bound the local memories and obtain $\mathbb{E}\|\frac{1}{R}\sum_{r=1}^{R} m_t^{(r)}\|^2 \le \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|m_t^{(r)}\|^2 \le 4C\frac{\eta_t^2}{\gamma^2}G^2H^2$. This implies

$$\mathbb{E}\|\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^{R}\widetilde{\mathbf{x}}_{t_r}^{(r)}\|^2 \le 4C\frac{\eta_t^2}{\gamma^2}G^2H^2. \tag{99}$$

Putting the bounds from (91), (94), and (99) in (90) gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \le 192B\left(1+\frac{C}{\gamma^2}\right)\eta_t^2H^4G^2 + 12C\frac{\eta_t^2}{\gamma^2}G^2H^2$$

This completes the proof of Lemma 13. □

C.6 Bounded distance between virtual and true sequence under fixed learning rate

Lemma 14 (Bounded Distance between Virtual and True Sequence)**.** *Let $\mathcal{I}_T^{(r)} \in [T]$ be a set of time instances in which the worker r updates and synchronizes with the master. For $\eta_t = \eta$, gap$(\mathcal{I}_T^{(r)} \le H)$ and $t \in \mathbb{Z}^+$ we have*

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \le 6C'\eta^2H^4G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2 \tag{100}$$

Here $C' = B\left(\frac{8}{\gamma^2}-6\right)$ where B is from $\mathbb{E}_{Q,C}\|QComp_k(\mathbf{x})\|^2 \le B\|\mathbf{x}\|^2$.

Proof. For a constant learning rate the first term in (90) has been bounded earlier in (86). Following similar steps as in (93) we would have

$$\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2 \le 2B\left(\frac{4}{\gamma^2}-3\right)\eta^2H^4G^2 \tag{101}$$

Finally using (86),(98), Lemma 3 and (101) in (90) we have

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \le 12B\left(\frac{4}{\gamma^2}-3\right)\eta^2H^4G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2 \tag{102}$$

□

D Further Experiments

D.1 Experiments for Convex Objective

The experiments in Figure 3-4 and in Figure 2 are in a synchronous distributed setting with 15 worker nodes, each processing a mini-batch size of 8 samples per iteration using the *MNIST* [19] handwritten digits dataset. The corresponding experiments for the asynchronous operation (as in Algorithm 2) are shown in Figure 5.

D.1.1 Model Architecture

Define the softmax function as

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

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

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

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

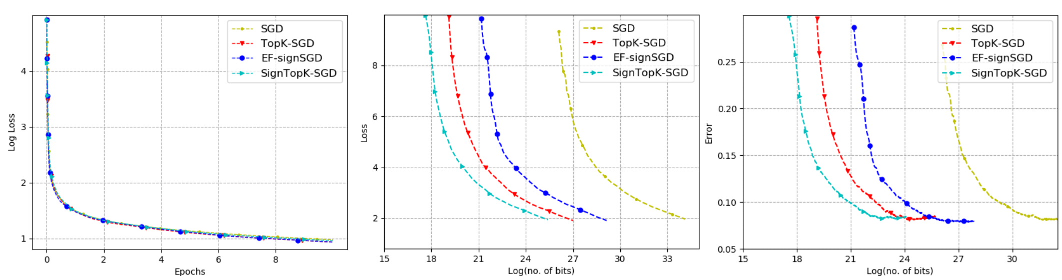
$$\mathbf{x} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_L], \quad \mathbf{x}_j \in \mathbb{R}^d, \ \forall j \in [L],$$

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

D.1.2 Parameter Selection and Learning Rates

We use our composed operator $SignTop_k$ (from Lemma 6) for compression, and denote the resulting SGD algorithm by SignTopK. The schemes with which we compare SignTopK are EF-SIGNSGD [15], TopK-SGD [4, 30], and local SGD [31]. 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 2, 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.

D.1.3 Results



(a) Training loss vs epochs

(b) Training loss vs $\log_2$ of communication budget

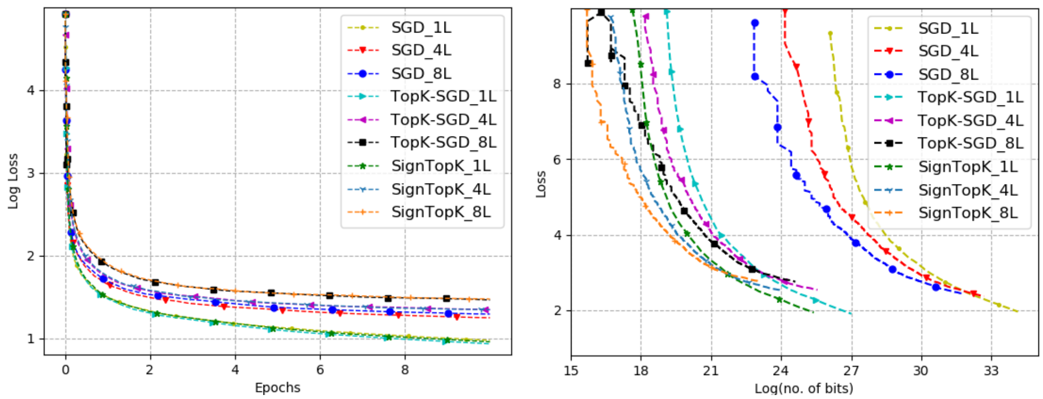
(c) top-1 accuracy [18] for schemes in Figure 3a

Figure 3 Figure 3a-3c demonstrate the gains in performance achieved by our $Qsparse$ operators in a convex setting.

In Figure 3a, 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. From Figure 3b and 3c, we see that our composed operator achieves gains in communicated bits by a factor of 6-8 times over the state-of-the-art.

Figure 4a, demonstrates the effect of incorporating local iterations in SignTopK, 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 h, SignTopK_hL maintains the same rates as vanilla SGD or TopK-SGD. In doing so, it is able to achieve gains in communicated bits as seen in Figure 4b, simply by communicating infrequently with the master.

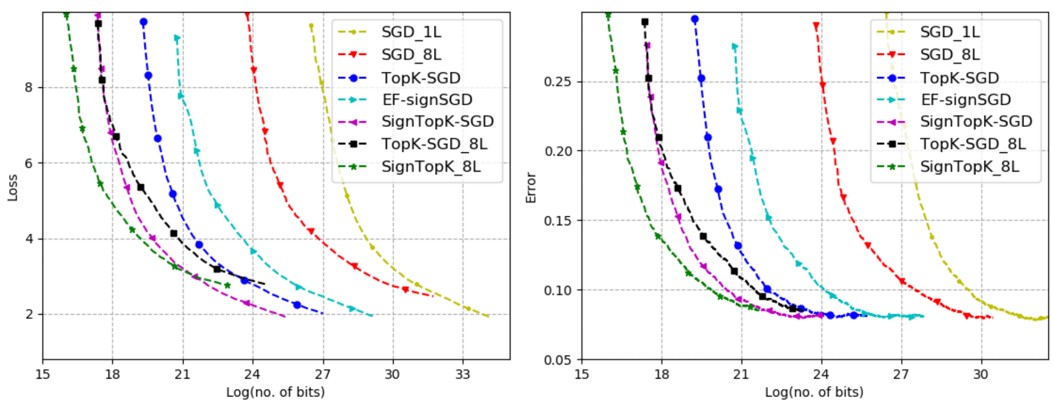
We observe similar trends in Figure 5a-5b for our asynchronous operation, where workers synchronize with the master at arbitrary time intervals as per Algorithm 2. Specifically, in our experiments, for each $r \in [R]$, the time interval for the rth worker is decided uniformly at random from $[H]$ after 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.



(a) Training loss vs epochs

(b) Training loss vs $\log_2$ of communication budget

Figure 4 Figure 4a-4b demonstrate the effect of incorporating local iterations and compare these effects across vanilla SGD, TopK-SGD, as well as SignTopK.



(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 5a

Figure 5 Figure 5a-5b demonstrate the performance of our scheme in comparison with EF-SIGNSGD [15] and TopK-SGD [4, 30] in a convex setting for asynchronous operation.

E Summary of Results

Combining local computations with quantization and explicit sparsification enables significantly reduced communication, resulting in a lot of bit savings. For a fixed number of local iterations H, we characterize the required total number of iterations $T = \Omega(\cdot)$ (see Table 1 and Table 2) after which the algorithm converges at the rates of distributed vanilla SGD. Furthermore, we also characterize the reduction in communication, in terms of the asymptotic limits of local computations, i.e., $H = \mathcal{O}(\cdot)$ (see Table 1 and Table 2).

		Synchronous	
Objective	Rate	H	T
Smooth and non-convex	$\mathcal{O}(1/\sqrt{bRT})$	$\mathcal{O}(\gamma T^{1/4}/(bR)^{3/4})$	$\Omega(H^4(bR)^3/\gamma^4)$
Smooth and strongly convex	$\mathcal{O}(1/bRT)$	$\mathcal{O}(\gamma\sqrt{T/(bR)})$	$\Omega(H^2(bR)/\gamma^2)$

Table 1 Summary of results for the synchronous setting with fixed learning rate in both the smooth and non-convex case and decaying learning rate in the smooth and strongly convex case.

		Asynchronous	
Objective	Rate	H	T
Smooth and non-convex	$\mathcal{O}(1/\sqrt{bRT})$	$\mathcal{O}(\sqrt{\gamma}T^{1/8}/(bR)^{3/8})$	$\Omega(H^8(bR)^3/\gamma^4)$
Smooth and strongly convex	$\mathcal{O}(1/bRT)$	$\mathcal{O}(\sqrt{\gamma}(T/(bR))^{1/4})$	$\Omega(H^4(bR)/\gamma^2)$

Table 2 Summary of results for the asynchronous setting with fixed learning rate in both the smooth and non-convex case and decaying learning rate in the smooth and strongly convex case.