Title

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

Author

Basu, Debraj

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

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

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Electrical and Computer Engineering

by

Debraj Debashish Basu

2019

ABSTRACT OF THE THESIS

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

by

Debraj Debashish Basu
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Suhas N. Diggavi, Chair

Large scale distributed optimization has become increasingly important with the emergence of edge computation architectures such as in the federated learning setup, where large amounts of data, possibly of a secure nature and generated in an online manner can be massively distributed across personal devices. A key bottleneck for many such large-scale problems is in the communication overhead of exchanging information between devices over bandwidth limited networks as well as in the unreliability of communication for distributed optimization. The existing approaches propose to mitigate these bottlenecks either by using different forms of compression or by computing local models and mixing them iteratively. In this thesis we first propose a novel class of highly communication efficient operators that employ stochastic and deterministic quantization with aggressive sparsification such as Top-k in the form of a composed operator. Furthermore, in federated learning one can use local computations to reduce communication. Using such a framework, we incorporate local iterations into our algorithm which allows the communication to be infrequent and possibly asynchronous thereby enabling significantly reduced communication.

Putting them together we have distributed *Qsparse-local-SGD* for federated learning for which our analysis demonstrates convergence rates matching vanilla distributed SGD where

we observe that quantization and sparsification are almost for "free" for smooth functions, both non-convex and convex. We characterize the asymptotic allowable limits of local iterations for synchronous and asynchronous implementations of *Qsparse-local-SGD*, so as to harness both the distributed processing gains as well as the benefits of quantization, sparsification and local computations. Our numerics demonstrate that *Qsparse-local-SGD* combines the bit savings of our composed operators, as well as local computations, thereby outperforming the cases where these techniques are individually used. We use it to train ResNet-50 on ImageNet, as well as a softmax multi-class classifier on MNIST, resulting in significant savings over the state-of-the-art, in the number of bits transmitted to reach target accuracy.

The thesis of Debraj Debashish Basu is approved.

Christina Panagio Fragouli

Lieven Vandenberghe

Suhas N. Diggavi, Committee Chair

University of California, Los Angeles

2019

To my parents, Areta and Debashis.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Related Work 2

1.2 Contributions 4

1.3 Organization 5

1.4 Preliminaries 6

1.4.1 Smoothness and Convexity 6

1.4.2 Useful Vector Inequalities 7

2 Communication Efficient Operators **8**

2.1 Quantization 8

2.2 Sparsification 10

2.3 Composed Operators 11

2.4 Discussion 14

3 Distributed Synchronous Operation **15**

3.1 Qsparse-local-SGD 15

3.2 Assumptions 17

3.3 Error Compensation 17

3.3.1 Decaying Learning Rate 18

3.3.2 Fixed Learning Rate 19

3.4 Main Results 19

3.5 Proof Outline 21

3.5.1 Proof outline of Theorem 1 21

3.5.2 Proof outline of Theorem 2 . . . 23
3.6 Discussion . . . 24

4 Distributed Asynchronous Operation . . . 25
4.1 Asynchronous Operation . . . 25
4.2 Main Results . . . 27
4.3 Proof Outline . . . 28
4.4 Discussion . . . 30

5 Communication Cost and Experiments . . . 31
5.1 Communication Cost . . . 31
5.2 Summary of Results . . . 33
5.3 Non Convex Objective . . . 34
5.3.1 Experiment setup . . . 34
5.3.2 Results . . . 34
5.4 Convex Objective . . . 35
5.4.1 Model Architecture . . . 36
5.4.2 Parameter selection and Learning rates . . . 36
5.4.3 Experiment Results . . . 37

6 Conclusion . . . 40

A Supplementary material for preliminaries in Chapter 1 . . . 41

B Supplementary material for Chapter 2 . . . 44
B.1 Proof of Lemma 1 . . . 44
B.2 Proof of Lemma 2 . . . 45

B.3 Proof of Lemma 3 47

C Supplementary material for Chapter 3 50

C.1 Proof of Lemma 4 50

C.2 Proof of Lemma 5 53

C.3 Proof of Lemma 6 54

C.4 Proof of Lemma 7 55

C.5 Proof of Lemma 8 56

C.6 Smooth Objective: Proof of Theorem 1 56

C.7 Convex Objective: Proof of Theorem 2 59

D Supplementary material for Chapter 4 64

D.1 Proof of Lemma 9 64

D.2 Proof of Lemma 10 66

D.3 Proof of Lemma 11 67

D.4 Proof of Lemma 12 70

E Supplementary material with additional results 71

E.1 Synchronous 71

E.2 Asynchronous 71

E.3 Proof of Theorem 5 72

References 74

LIST OF FIGURES

5.1 Figures 5.1a-5.1c demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19], TopK-SGD [SCJ18, AHJ18] and local SGD [Sti19, YYZ18] in a non convex setting. 35

5.2 Figures 5.2a-5.2c demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19] and TopK-SGD [SCJ18, AHJ18] in a convex setting for synchronous updates. Here for $H = 1, 4, 8, 16$, corresponds to the Algorithm 1 running with a synchronization period of at most H. 37

5.3 Figures 5.3a-5.3b demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19] and TopK-SGD [SCJ18, AHJ18] in a convex setting for asynchronous operation. 39

LIST OF TABLES

5.1 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. 33

5.2 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. 33

ACKNOWLEDGMENTS

My experience as a Masters student at UCLA has been exceedingly enriching, for which I am most grateful to my advisor, Professor Suhas Diggavi. I would like to thank him for his patience, generosity and unparalleled guidance over this duration. None of this would have been possible without his encouragement and support in many forms, for which I shall always remain indebted to him. Much of the contents of this thesis are a result of the foundations which were laid in classes offered by members of my thesis committee, Professor Diggavi, Professor Vandenberghe and Professor Fragouli, all of which truly inspired me and I thoroughly enjoyed them, both as a student, as well as a teaching assistant.

I have also had the privilege of enjoying the intellectual company of Dr. Deepesh Data and Dr. Can Karakus among many others, whose role has been critical in shaping this thesis. It has been an incredible journey undertaking research in collaboration with Deepesh, who has been a great mentor and friend, and whose expertise in several domains facilitated both our learning processes, eventually culminating in this thesis. I am also much obliged to Can, who not only introduced me to this area of research at the intersection of communication theory and machine learning, but always encouraged me to identify the right problem, and to make mistakes and learn from them. A special mention for Can, who has very kindly invested a lot of time around his other professional commitments without which this thesis would not have been possible.

I must admit that I have been very lucky throughout this journey. I would like to thank UCLA for the amazing curriculum that it offers which perfectly aligned with my interests in optimization, information theory and automated reasoning. As a teaching assistant, I have also had the opportunity of interacting with members of the faculty, as well as numerous graduate and undergraduate students, with different backgrounds. This diverse interaction, provided me with different perspectives for analyzing problems, and further enhanced my learning process. This is probably one of the things that I am going to miss the most about my time spent at UCLA.

I would like to also acknowledge several insightful discussions with Navjot in the early stages of this work, as well as the lunch and coffee breaks with Dhaivat, Navjot and Deepesh, which I would keep looking forward to each day. I must also mention all my labmates and friends, Deepesh, Mehrdad, Dhaivat, Navjot, Kenny, Sundar, Antonious, Osama, Yahya, Gaurav and Karmoose with whom I have had academic and friendly interactions and I wish them the very best for their current and future endeavors.

I would like to thank Sakshi, for her support, encouragement, and for being an amazing girlfriend. I am very proud of her. And finally I would like to thank my parents, Debashis and Areta who have been constant pillars of support throughout my life. Their immeasurable love and wisdom has been the driving force in my decisions and has always guided me well.

CHAPTER 1

Introduction

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

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

where $\{g_t^r\}_{r=1}^R$ are the local stochastic gradients at the R machines for a local loss function $f^{(r)}(\mathbf{x})$ of the parameters, where $f^{(r)} : \mathbb{R}^d \to \mathbb{R}$ and η_t is the learning rate.

The training of high dimensional models is done at a large scale over bandwidth limited networks. Therefore despite the distributed processing gains, it is well understood by now that exchange of full-precision gradients between nodes, causes communication to be the bottleneck for many large scale models [AHJ18,WXY17,BWA18,SYK17]. For example, consider training the ResNet 152 architecture [HZR16] which has about 60 million parameters, on the ImageNet dataset that contains 14 million images. Each full precision exchange between workers is around 0.24 GB. And this problem is at a much smaller scale in comparison to the real world platforms where data resides on personal devices such as laptops and tablets. Therefore the communication bottleneck could be significant in emerging edge computation architectures suggested by federated learning [Kon17, MMR17, ABC16]. To address this, many methods have been proposed recently, and these methods are broadly based on three major approaches:

1. *Quantization* of gradients, where nodes locally quantize the gradient (perhaps with randomization) to a small number of bits [AGL17, BWA18, WHH18, WXY17, SYK17].

2. *Sparsification* of gradients, *e.g.,* where nodes locally select Top_k values of the gradient in absolute value and transmit these at full precision [Str15, AH17, SCJ18, AHJ18, WHH18, LHM18], while maintaining errors in local nodes for later compensation.

3. *Skipping communication rounds* whereby nodes average their models after locally updating their models for several steps [YYZ18, Cop15, ZDW13, Sti19, CH16, WJ18].

In this work we propose *Qsparse-local-SGD* algorithm, which combines aggressive sparsification with quantization and local computations, along with error compensation, by keeping track of the difference between the true and compressed gradients. We propose both synchronous and asynchronous implementations of *Qsparse-local-SGD* in a *distributed* setting where the nodes perform computations on their local datasets. In our asynchronous model, the distributed nodes' iterates evolve at the same rate, but update the gradients at arbitrary times; see Chapter 4 for more details. 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 for a softmax multiclass classifier using the MNIST dataset, and show that we achieve target accuracies with about a factor of 15-20 savings over the state-of-the-art [AHJ18, SCJ18, Sti19], in the total number of bits transmitted.

1.1 Related Work

The use of quantization for communication efficient gradient methods has decades rich history [GMT73] and its recent use in training deep neural networks [SFD14, Str15] has re-ignited interest. Theoretically justified gradient compression using unbiased stochastic quantizers has been proposed and analyzed in [AGL17, WXY17, SYK17]. Though methods

in [WWL18, WSL18] use induced sparsity in the quantized gradients, explicitly sparsifying the gradients more aggressively by retaining Top_k components, *e.g.,* $k < 1\%$, has been proposed [Str15, AH17, LHM18, AHJ18, SCJ18], combined with error compensation to ensure that all co-ordinates do get eventually updated as needed. [WHH18] 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*, see [Cop15], and *model averaging*, see [Sti19, YYZ18, ZSM16] and references therein. With the onset of powerful processing units such as graphics and tensor processing units, increasing the computation to communication ratio would enable distributed algorithms to converge faster. Furthermore, this would also be useful in applications where compression does not result in significant gains, such as in cloud computing based frameworks where the data resides online and the computations are also performed online. Our work is most closely related to and builds on the recent theoretical results in [AHJ18, SCJ18, Sti19, YYZ18]. [SCJ18] considered the analysis for the centralized Top_k (among other sparsifiers), and [AHJ18] 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 [AHJ18]. [Sti19, YYZ18] 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 [AHJ18], by using the perturbed iterate methods [MPP17, SCJ18]. We analyze non-convex (smooth) objectives as well as strongly convex objectives for the distributed case with local computations. [SCJ18] gave a proof of sparsified SGD, for convex objective functions and for centralized case, without local computations. Our techniques compose a (stochastic or deterministic 1-bit sign) quantizer with sparsification and local computations using error compensation. While our focus has only been on mitigating the communication bottlenecks in training high dimensional models over bandwidth limited networks, this technique works for any compression operator satisfying a regularity condition

(see Definition 7) including our composed operators. Operators satisfying this condition are becoming increasingly popular in several recent works such as in [KRS19, KSJ19], and our composed operators directly find application in such settings.

1.2 Contributions

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

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

where $f^{(r)}(\mathbf{x}) = \mathop{\mathbb{E}}_{i\sim\mathcal{D}_r}[f_i(\mathbf{x})]$, where $\mathop{\mathbb{E}}_{i\sim\mathcal{D}_r}[\cdot]$ denotes expectation over a random sample chosen from the local data set $\mathcal{D}_r$. Our setup can also handle different local functional forms, beyond dependence on the local data set $\mathcal{D}_r$, which is not explicitly written for notational simplicity. For $f : \mathbb{R}^d \to \mathbb{R}$, 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. With some minor modifications to *Qsparse-local-SGD*, it can also be used in a peer to peer setting where the aggregation is done without any help from a master node, and each worker exchanges its updates with all other workers.

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 opera-

tion affords, and has convergence similar to equivalent vanilla SGD with local iterations (see Corollary 2, 3), for both the non-convex case (with convergence rate $\sim \frac{1}{\sqrt{T}}$ for fixed learning rate) as well as the strongly convex case (with convergence rate $\sim \frac{1}{T}$, for diminishing learning rate); demonstrating that quantizing and sparsifying the gradient, even after local iterations asymptotically yields an almost "free" efficiency gain (also observed numerically in Chapter 5 non-asymptotically). The numerical results on ImageNet dataset implemented for a ResNet-50 architecture and for the convex case for multi-class logistic classification on MNIST [LBB98] dataset demonstrates that one can get significant communication savings, while retaining equivalent state-of-the art performance. The combination of quantization, sparsification and local computations poses several challenges for theoretical analysis, including the analysis of impact of local iterations (block updates) of parameters on quantization and sparsification (see Lemma 4-5 in Chapter 3); as well as asynchronous updates and its combination with distributed compression (see Lemma 9-12 in Chapter 4).

1.3 Organization

Chapter 2 introduces a novel class of operators that help mitigate the communication bottlenecks in distributed optimization. Distributed SGD when run in combination with such operators motivates a highly communication efficient algorithm for large scale distributed optimization that combines quantization, sparsification and local computations in Chapter 3. Chapter 3 and Chapter 4 outline the main technical results and skeleton of the proofs for both non-convex and convex functions. Finally, Chapter 5 compares the gains in communication achieved by *Qsparse-local-SGD* and numerically demonstrates the performance of the schemes for both non-convex and convex distributed optimization. A short discussion in Chapter 6 concludes the thesis.

1.4 Preliminaries

The scope of this work is limited to functions that are continuously differentiable and unless explicitly mentioned, $\|\cdot\|$ is used to denote the euclidean norm $\|\cdot\|_2$. The convergence analyses of our algorithm *Qsparse-local-SGD*, is from first principles with the basic analytical tools for non-convex and convex optimization provided below. The corresponding explanations for this toolset are provided in Appendix A.

Definition 1 (Continuously Differentiable Function)**.** *A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is a function whose gradient denoted by $\nabla f(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \mathbb{R}^d$.*

1.4.1 Smoothness and Convexity

Definition 2 (Convex Function)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if Jensen's inequality holds i.e. for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \quad \forall \lambda \in [0,1].$$

Remark 1. *For continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$, a first order condition for convexity is that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Definition 3 (Smooth Function)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth for parameter $L \geq 0$ if the gradients are Lipschitz continuous, i.e., for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ the following holds*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Together with the Taylor expansion, the following holds

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Remark 2. *If a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex with minimizer $\boldsymbol{x}^*$ such that $\nabla f(\boldsymbol{x}^*) = 0$, then*

$$\|\nabla f(\boldsymbol{x})\|_2^2 = \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*)\|_2^2 \leq 2L(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)).$$

Definition 4 (Strongly Convex Function)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for parameter $\mu \geq 0$ if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ the following holds*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

Remark 3. *If a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with minimizer $\boldsymbol{x}^*$ such that $\nabla f(\boldsymbol{x}^*) = 0$, then*

$$\|\nabla f(\boldsymbol{x})\|_2^2 = \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*)\|_2^2 \geq 2\mu(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)).$$

1.4.2 Useful Vector Inequalities

Remark 4. *For n arbitrary vectors $\{\boldsymbol{u}_i\}_{i=1}^n, \boldsymbol{u}_i \in \mathbb{R}^d$, the following holds by the convexity of $\|\cdot\|_2$ and Jensen's inequality,*

$$\left\| \sum_{i=1}^n \boldsymbol{u}_i \right\|_2^2 \leq n \sum_{i=1}^n \|\boldsymbol{u}_i\|_2^2.$$

Remark 5. *For n arbitrary vectors $\{\boldsymbol{u}_i\}_{i=1}^n, \boldsymbol{u}_i \in \mathbb{R}^d$, with $\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i$, the following holds,*

$$\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{u}_i - \overline{\boldsymbol{u}}\|_2^2 = \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{u}_i\|_2^2 - \|\overline{\boldsymbol{u}}\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{u}_i\|_2^2.$$

Remark 6. *For any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\gamma > 0$, the following holds,*

$$2\langle \boldsymbol{u}, \boldsymbol{v} \rangle \leq \gamma \|\boldsymbol{u}\|_2^2 + \gamma^{-1} \|\boldsymbol{v}\|_2^2.$$

As a consequence we will also have $\|\mathbf{u} + \mathbf{v}\|_2^2 \leq (1 + \gamma)\|\mathbf{u}\|_2^2 + (1 + \gamma^{-1})\|\mathbf{v}\|_2^2$

CHAPTER 2

Communication Efficient Operators

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

We first consider two different techniques used in the literature for mitigating the communication bottleneck in distributed optimization, namely, quantization and sparsification. In quantization, we reduce precision of the gradient vector by mapping each of its components using a deterministic [BWA18, KRS19] or randomized [AGL17, WXY17, SYK17, ZDJ13] 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, [SCJ18, KSJ19].

2.1 Quantization

SGD computes an unbiased estimate of the gradient which can be used to update the model iteratively and is extremely useful in large scale applications. It is well known that the first

order terms in the rate of convergence are affected by the variance of the gradients which have an upper bound of σ^2. While stochastic quantization of gradients could result in a variance blow up, it preserves the unbiasedness of the gradients at low precision, and therefore when training over bandwidth limited networks, the convergence would be much faster.

Definition 5 (Randomized Quantizer [AGL17, WXY17, SYK17, ZDJ13])**.** *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)* $\mathbb{E}_Q[\|Q_s(\mathbf{x})\|^2] \leq (1 + \beta_{d,s})\|\mathbf{x}\|^2$*, where* $\beta_{d,s} > 0$ *could be a function of* d *and* s*. Here expectation is taken over the randomness of* Q_s*.*

Examples of randomized quantizers include

1. *QSGD* [AGL17, WXY17], which independently quantizes components of $\mathbf{x} \in \mathbb{R}^d$ into s levels, with $\beta_{d,s} = \min(\frac{d}{s^2}, \frac{\sqrt{d}}{s})$);
2. *Stochastic* s*-level Quantization* [SYK17, ZDJ13], 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 $\beta_{d,s} = \frac{d}{2s^2}$;
3. *Stochastic Rotated Quantization* [SYK17], 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 [BWA18, KRS19].

Definition 6 (Deterministic Sign Quantizer [BWA18, KRS19])**.** *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* $\mathbb{1}\{x_i \geq 0\} - \mathbb{1}\{x_i < 0\}$*.*

Such methods drew interest since RPROP [RB93] which only used the temporal behavior of the sign of the gradient. This is an example where the biased 1-bit quantizer as in Definition 6 is used. This further inspired optimizers such as RMSPROP [TH12], and ADAM [KB15], which incorporate appropriate adaptive scaling with momentum acceleration and have demonstrated empirical superiority in non-convex applications.

2.2 Sparsification

As mentioned earlier, we consider two important examples of sparsification operators: Top_k and Rand_k, For any $\mathbf{x} \in \mathbb{R}^d$, $\text{Top}_k(\mathbf{x})$ 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, $\text{Rand}_k(\mathbf{x})$ 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 "contraction" property as defined below, with $\gamma = k/d$ [SCJ18]. Few other examples of such operators can be found in [KRS19, KSJ19].

Definition 7 (Contraction Operator [SCJ18])**.** *A (randomized) function $Comp_k : \mathbb{R}^d \to \mathbb{R}^d$ is called a* contraction *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 $Comp_k$.*

Note that stochastic quantizers as in Definition 5 also satisfy this regularity condition in Definition 7 for $\beta_{d,s} \leq 1$. Now we give a simple but important corollary, which allows us to apply different contraction operators to different coordinates of a vector. For example, in the case of training neural networks, we can apply different operators to different layers.

Corollary 1 (Piecewise Contraction)**.** *Let $C_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ for $i \in [L]$ denote possibly different contraction operators with contraction 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 $C(\boldsymbol{x}) := [C_1(\mathbf{x}_1)\, C_2(\mathbf{x}_2)\, \ldots C_L(\mathbf{x}_L)]$ is a contraction operator with the contraction 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 contraction operator with the contraction coefficient γ_i. □

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

2.3 Composed Operators

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

Lemma 1 (Contraction of the Composed Operator)**.** *Let $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ be a quantizer with parameter s that satisfies Definition 5. 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 $\mathbf{x} \in \mathbb{R}^d$. If k, s are such that $\beta_{k,s} < 1$. then $Q_sComp_k : \mathbb{R}^d \to \mathbb{R}^d$ is a contraction operator with the contraction coefficient being equal to $\gamma = (1 - \beta_{k,s})\frac{k}{d}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, we have*

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

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

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

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

Observe that for a given stochastic quantizer that satisfies Definition 5, we have a prescribed operating regime of $\beta_{k,s} < 1$. This results in an upper bound on the coarseness of the quantizer, which happens because the quantization leads to a blow-up of the second moment; see condition (ii) of Definition 5. However, by employing Corollary 1, we show that this can be alleviated to some extent via an example.

Remark 7. *Consider an operator as described in Lemma 1, where the quantizer, $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ in use is QSGD [AGL17, WXY17], and the sparsifier, $Comp_k$ is Top_k [AHJ18, SCJ18]. Apply it to a vector $\mathbf{x} = [\mathbf{x}_1\, \mathbf{x}_2\, \dots\, \mathbf{x}_L] \in \mathbb{R}^d$ in a piecewise manner, i.e., $Q_{s_i}Comp_{k_i} : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ to smaller vectors $\mathbf{x}_i \in \mathbb{R}^{d_i}$ as prescribed in Corollary 1. Define $\beta_{k_i,s_i} = \frac{k_i}{s_i^2}$ as the coefficient of the variance bound as in Definition 5 for the quantizer Q_{s_i}, used for $\mathbf{x}_i$ and $k := \sum_{i=1}^{L} k_i$. Observe that the regularity condition in Definition 7 can be satisfied by having $k_i < s_i^2$. Therefore the piecewise contraction operator allows a coarser quantizer than when the operator is applied to the entire vector together, where we require $\beta_{k,s} = \frac{k}{s^2} < 1$, thus providing a small gain in communication efficiency. For example, consider the composed operator being applied on a per layer basis to a deep neural network. We can now afford to have a much coarser quantizer than when the operator is applied to all parameters at once.*

As discussed earlier, stochastic quantization results in a variance blow-up which limits our regime of operation. However we can scale the quantized vector, $Q_sComp_k(\mathbf{x})$, appropriately as presented in Lemma 2, so as to mitigate the variance blow up. A proof is provided in Appendix B.2

Lemma 2 (Composing sparsification with stochastic quantization)**.** *Let operator $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ be a stochastic quantizer with parameter s that satisfies Definition 5. 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 contraction operator with the contraction coefficient being equal to $\gamma = \frac{k}{d(1+\beta_{k,s})}$, i.e., for every $\boldsymbol{x} \in \mathbb{R}^d$*

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

Remark 8. *Note that, unlike* $Q_s Comp_k(\mathbf{x})$*, the* scaled *version* $\frac{Q_s Comp_k(\mathbf{x})}{1+\beta_{k,s}}$ *is always a contraction operator for all values of* $\beta_{k,s} > 0$*. Furthermore, observe that, if* $\beta_{k,s} < 1$*, then we have* $(1-\beta_{k,s})\frac{k}{d} < \frac{k}{d(1+\beta_{k,s})}$*, which implies that even in the operating regime of* $\beta_{k,s} < 1$*, which is required in Lemma 1, the* scaled *composed operator* $\frac{Q_s Comp_k(\mathbf{x})}{1+\beta_{k,s}}$ *of Lemma 2 gives better contraction than what we get from the* unscaled *composed operator* $Q_s Comp_k(\mathbf{x})$ *of Lemma 1. So, scaling a composed operator properly is always a better choice for contraction.*

We can also compose a *deterministic* quantizer with $Comp_k$. For that we need some notations first. For $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$ and given vector $\mathbf{x} \in \mathbb{R}^d$, let $\mathcal{S}_{Comp_k(\mathbf{x})} \in \binom{[d]}{k}$ denote the set of k indices chosen for defining $Comp_k(\mathbf{x})$. For example, if $Comp_k = \mathrm{Top}_k$, then $\mathcal{S}_{Comp_k(\mathbf{x})}$ denote the set of k indices corresponding to the largest component of $\mathbf{x}$; if $Comp_k = \mathrm{Rand}_k$, then $\mathcal{S}_{Comp_k(\mathbf{x})}$ denote a set of random set of k indices in $[d]$. For any $x \in \mathbb{R}$ we define $Sign(x) := \mathbb{1}\{x \geq 0\} - \mathbb{1}\{x < 0\}$ as in Definition 6. The composition of $Sign$ with $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$ is denoted by $SignComp_k : \mathbb{R}^d \to \mathbb{R}^d$, and for $i \in [d]$, the i'th component of $SignComp_k(\mathbf{x})$ is defined as

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

Now we show that the composition of $Sign$ with $Comp_k \in \{Top_k, Rand_k\}$ is a contraction operator. A proof is provided in Appendix B.3.

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

$$\frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k}$$

for any $m \in \mathbb{Z}_+$ *is a contraction operator with the contraction 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 9. *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.*

2.4 Discussion

In Chapter 2 we propose a novel class of composed operators that combine the bit savings of existing powerful techniques for compression. Operators satisfying such regularity conditions have shown promise in [SCJ18, KRS19, KSJ19] and our composed operators directly find application in such settings. In fact, in Chapter 3 and 4 we lay the foundations of our algorithm for empirical risk minimization over distributed nodes, *Qsparse-local-SGD* that further incorporates infrequent and possibly asynchronous communication which when used in combination with our composed operators results in significant savings in the number of bits exchanged while converging at rates matching vanilla distributed SGD.

CHAPTER 3

Distributed Synchronous Operation

Chapter 2 combines stochastic and deterministic quantization with highly aggressive sparsification. Chapter 1 refers to literature on local computations in which workers perform multiple local iterations between two round of communication. In a distributed setup where communication is the major bottleneck, skipping communication rounds would significantly mitigate this bottleneck. What remains to be seen is whether we can achieve this at rates matching vanilla SGD as well.

Section 3.1 provides an algorithm that combines three different techniques of improving the communication efficiency of distributed SGD. Namely, *Qsparse-local-SGD* combines quantization, sparsification and local computations to result in fast convergence at a highly reduced communication cost. The updates are performed in a synchronized manner and that is further relaxed in Section 4.1 where every worker maintains a local sequence as earlier but has its own update schedule. Both in the synchronous and asynchronous setup each worker performs a finite number of local iterations before synchronizing its parameter with the master. These algorithms are useful for performing empirical risk minimization both in the case of non-convex and convex loss functions.

3.1 Qsparse-local-SGD

Let $\mathcal{I}_T^{(r)} \subseteq [T] := \{1, \ldots, T\}$ 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. 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.

Algorithm 1 Qsparse-local-SGD

Initialize $\mathbf{x}_0 = \widehat{\mathbf{x}}_0^{(r)} = m_0^{(r)} = \mathbf{0},\ \forall r \in [R]$. 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 $t+1 \notin \mathcal{I}_T$ **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
      $g_t^{(r)} \leftarrow Q\,Comp_k\left(m_t^{(r)} + \mathbf{x}_t - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}\right)$, 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: $\widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}$ is used to denote an intermediate variable between iterations t and $t+1$.

3.2 Assumptions

The analysis follows in the following Chapters and we make the following standard assumptions with references to prior art.

1. **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 $f^{(r)}(\mathbf{y}) \leq f^{(r)}(\mathbf{x}) + \langle \nabla f^{(r)}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$.

2. **Bounded second moment:** For every $\widehat{\mathbf{x}}_t^{(r)} \in \mathbb{R}^d, r \in [R], t \in [T]$ and for some constant $0 \leq G < \infty$, we have $\mathop{\mathbb{E}}_{i \sim \mathcal{D}_r}[\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)})\|_2^2] \leq G^2$. This is a standard assumption in [SSS07, NJL09, RRW11, HK14, RSS12, SCJ18, Sti19, YYZ18, KSJ19, AHJ18]. Relaxation of the uniform boundedness of the gradient allowing arbitrarily different gradients of local functions in heterogenous settings as done for SGD in [NND18, WJ18] is left as future work. This also imposes a **bound on the variance**: $\mathop{\mathbb{E}}_{i \sim \mathcal{D}_r}[\|\nabla f_i(\widehat{\mathbf{x}}_t^{(r)}) - \nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|_2^2] \leq \sigma_r^2$, where $\sigma_r^2 \leq G^2$ for every $r \in [R]$.

In this section we present our main convergence results with synchronous updates, obtained by running the Algorithm 1 for smooth functions, both non-convex and strongly convex. To state our results, we need the following definition from [Sti19].

Definition 8 (Gap [Sti19])**.** *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.*

3.3 Error Compensation

Sparsified gradient methods where the workers send the top-k coordinates of the updates based on their magnitudes have been investigated in the literature, and serves as a communication efficient strategy for distributed training of learning models. However the convergence rates are subpar to distributed vanilla SGD. Together with some form of error compensation, these methods have been empirically observed to converge as fast as vanilla

SGD in [Str15, AH17, LHM18, AHJ18, SCJ18]. In [AHJ18, SCJ18], sparsified SGD with such feedback schemes has been carefully analyzed. Under analytic assumptions [AHJ18], proves the convergence of parallel Top_k SGD with error feedback. The net error in the system is accumulated by each worker locally on a per iteration basis and this is used as feedback for generating the future updates. [SCJ18] did the analysis for the centralized Top_k SGD for convex objectives.

In algorithm 1 the historical error is accumulated into the memory of each worker which is compensated for in the future rounds of communication. This feedback is the key to recovering to convergence rates matching vanilla SGD. The operators employed provide a controlled way of using both the current update as well as the compression errors from the previous rounds of communication. Under the assumption of the uniform boundedness of the gradient we analyze the controlled evolution of memory through the optimization process in Lemma 4 and Lemma 5.

3.3.1 Decaying Learning Rate

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

$$\mathbb{E}\|m_t^{(r)}\|_2^2 \leq 4\frac{\eta_t^2}{\gamma^2} C H^2 G^2. \tag{3.1}$$

Therefore we see that the memory decays as $\mathcal{O}(\eta_t^2)$.This is a result of the "contraction" property Definition 7, and a proof is provided in Appendix C.1. This implies that the net error in the algorithm from the compression of updates in each round of communication is compensated for in the end. Similarly, as a result of the regularity condition we also have a bound for when the learning rate is fixed. The corresponding proof is provided in Appendix C.2

3.3.2 Fixed Learning Rate

Lemma 5 (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*

$$\mathbb{E}\|m_t^{(r)}\|_2^2 \leq 4\frac{\eta^2(1-\gamma^2)}{\gamma^2}H^2G^2. \tag{3.2}$$

For a fixed learning rate we observe that the memory is upper bounded by a constant $\mathcal{O}(\eta^2)$. Since the memory captures the historical errors due to compression, in order to asymptotically reduce the error to zero, the learning rate would have to be reduced once in a while throughout the training process.

3.4 Main Results

We leverage the perturbed iterate analysis as in [MPP17,SCJ18] to provide convergence guarantees for *Qsparse-local-SGD*. Under assumptions (i) and (ii), the following theorems hold when Algorithm 1 is run with any contraction 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 contraction operator whose contraction coefficient is equal to $\gamma \in (0, 1]$. Let $\{\widehat{\boldsymbol{x}}_t^{(r)}\}_{t=0}^{T-1}$ 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*

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \left(\frac{\mathbb{E}[f(\boldsymbol{x}_0)] - f^*}{\widehat{C}} + \widehat{C}L\left(\frac{\sum_{r=1}^R \sigma_r^2}{bR^2}\right)\right)\frac{4}{\sqrt{T}} + 8\left(4\frac{(1-\gamma^2)}{\gamma^2} + 1\right)\frac{\widehat{C}^2L^2G^2H^2}{T}. \tag{3.3}$$

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

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. With this in mind, see Corollary 2 below.

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

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \mathcal{O}\left(\frac{J\sigma_{max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2 bRG^2H^2}{\sigma_{max}^2\gamma^2 T}\right). \tag{3.4}$$

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

Here we characterize the reduction in communication that can be afforded; however, for a constant H, we get the same rate of convergence after $T = \Omega\left((bR)^3/\gamma^4\right)$ iterations. Analogous statements hold for Theorem 2-4 and are summarized in Section 5.2.

Theorem 1 is proved in Appendix C and provides non-asymptotic guarantees, where we observe that compression does not affect the first order term. Here we must decide the horizon T before running the algorithm, therefore for converging to a fixed point the learning rate needs to follow a piecewise schedule, which is also the case in our numerics in Section 5.3. 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 E.

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 contraction operator whose contraction coefficient is equal to* $\gamma \in (0, 1]$*. Let* $\{\widehat{\boldsymbol{x}}_t^{(r)}\}_{t=0}^{T-1}$ *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* $a > \max\{{}^{4H}/_{\gamma}, 32\kappa, H\}$*,* $\kappa = {}^{L}/_{\mu}$*. Then the following holds*

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \frac{La^3}{4S_T}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{8LT(T+2a)}{\mu^2 S_T}A + \frac{128LT}{\mu^3 S_T}B. \tag{3.5}$$

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

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \mathcal{O}\left(\frac{G^2H^3}{\mu^2\gamma^3T^3}\right) + \mathcal{O}\left(\frac{\sigma_{max}^2}{\mu^2 bRT} + \frac{H\sigma_{max}^2}{\mu^2 bR\gamma T^2}\right) + \mathcal{O}\left(\frac{G^2H^2}{\mu^3\gamma^2T^2}\right). \tag{3.6}$$

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

1. We recover [YYZ18, Theorem 1], which does local SGD for the non-convex case;
2. We generalize [Sti19, Theorem 2.2], which does local SGD for a strongly convex case and requires that each worker has identical datasets, to the distributed case.

3.5 Proof Outline

Maintain virtual sequences for every worker

$$\widetilde{\mathbf{x}}_0^{(r)} := \widehat{\mathbf{x}}_0^{(r)} \qquad \text{and} \qquad \widetilde{\mathbf{x}}_{t+1}^{(r)} := \widetilde{\mathbf{x}}_t^{(r)} - \eta_t \nabla f_{i_t^{(r)}}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \tag{3.7}$$

Define

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

3.5.1 Proof outline of Theorem 1

Proof. Since f is L-smooth we have from (3.7) that

$$f(\widetilde{\mathbf{x}}_{t+1}) - f(\widetilde{\mathbf{x}}_t) \leq -\eta_t \langle \nabla f(\widetilde{\mathbf{x}}_t), \mathbf{p}_t \rangle + \frac{\eta_t^2 L}{2}\|\mathbf{p}_t\|_2^2. \tag{3.8}$$

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

$$\frac{\eta_t}{4R}\sum_{r=1}^R \mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \eta_t^2 L \mathbb{E}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|_2^2 + 2\eta_t L^2 \mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|_2^2 + 2\eta_t L^2 \frac{1}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2. \tag{3.9}$$

Under Assumptions 1 and 2 provided in Section 3.2 we have

$$\mathbb{E}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|_2^2 \leq \frac{\sum_{r=1}^R \sigma_r^2}{R^2}. \tag{3.10}$$

To bound the third term on the RHS of (3.9) we first prove Lemma 6 following which we can use memory bounds in Section 3.3 to bound it.

Lemma 6 (Memory)**.** *The memory is maintained so as to capture the distance between the true sequence and virtual sequence.*

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^{R} m_t^{(r)}. \tag{3.11}$$

In Lemma 6 we show that the difference of the true and the virtual parameter vectors is equal to the average memory. A proof of Lemma 6 is provided in Appendix C.3. Using Lemma 5, we get

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

The fourth term on the RHS of (3.9) depicts the deviation of the local sequences which can be bounded as shown in Lemma 7. The details are provided in Appendix C.4.

Lemma 7 (Bounded Deviation of Local Sequences)**.** *With $\eta_t = \eta$ we have the following bound*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq \eta^2G^2H^2. \tag{3.12}$$

Using (3.10), Lemma 6, Lemma 5 and Lemma 7 in (3.9) we get

$$\frac{\eta}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta^2L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 8\frac{\eta^3(1-\gamma^2)}{\gamma^2}L^2G^2H^2 + 2\eta^3L^2G^2H^2. \tag{3.13}$$

Finally performing a telescopic sum from $t = 0$ to $T-1$ we get

$$\frac{1}{RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|_2^2 \leq \frac{4\left(\mathbb{E}[f(\widetilde{\mathbf{x}}_0)] - f^*\right)}{\eta T} + \frac{4\eta L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 32\frac{\eta^2(1-\gamma^2)}{\gamma^2}L^2G^2H^2 + 8\eta^2L^2G^2H^2. \tag{3.14}$$

For $\eta = \widehat{C}/\sqrt{T}$ we arrive at the statement of Theorem 1. □

3.5.2 Proof outline of Theorem 2

Proof. Using the definition of virtual sequences we have

$$\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|_2^2 + \eta_t^2 \|\mathbf{p}_t - \overline{\mathbf{p}}_t\|_2^2 - 2\eta_t \left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t, \mathbf{p}_t - \overline{\mathbf{p}}_t \right\rangle \tag{3.15}$$

On taking expectation with respect to the sampling at time t the third term on the RHS vanishes. With some algebraic manipulations provided in Appendix C.7, using the μ-strong convexity and L-smoothness of f for $\eta_t \leq 1/4L$ and letting $e_t = \mathbb{E}[f(\widehat{\mathbf{x}}_t)] - f^*$, we arrive at

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 \leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|_2^2 - \frac{\eta_t \mu}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) \mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_2^2 \\ + \frac{3\eta_t L}{R} \sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}.\end{aligned} \tag{3.16}$$

The fourth term on the RHS of (3.16) is the deviation of local sequences and can be bounded as in Lemma 8 for decaying learning rates. The details are provided in Appendix C.5.

Lemma 8 (Contracting Deviation of Local Sequences)**.** *Similar to Lemma 3.3 in [Sti19] we bound the deviation of the local sequences.*

$$\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|_2^2 \leq 4\eta_t^2 G^2 H^2. \tag{3.17}$$

Using (3.10), Lemma 6, Lemma 4 and Lemma 8 in (3.16) we get

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|_2^2 \leq \left(1 - \frac{\mu\eta_t}{2}\right) \mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|_2^2 - \frac{\mu\eta_t}{2L} e_t + \eta_t \left(\frac{3\mu}{2} + 3L\right) C \frac{4\eta_t^2}{\gamma^2} G^2 H^2 \\ + (3\eta_t L) 4\eta_t^2 L G^2 H^2 + \eta_t^2 \frac{\sum_{r=1}^R \sigma_r^2}{bR^2}.\end{aligned} \tag{3.18}$$

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

$$a_{t+1} \leq \left(1 - \frac{\mu\eta_t}{2}\right) a_t - \frac{\mu\eta_t}{2L} e_t + \eta_t^2 A + \eta_t^3 B. \tag{3.19}$$

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

$$\frac{\mu}{2LS_T} \sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu a^3}{8S_T} a_0 + \frac{4T\left(T + 2a\right)}{\mu S_T} A + \frac{64T}{\mu^2 S_T} B. \tag{3.20}$$

From convexity we can finally write

$$\mathbb{E}f\left(\overline{\mathbf{x}}_T\right) - f^* \leq \frac{La^3}{4S_T}a_0 + \frac{8LT\left(T+2a\right)}{\mu^2 S_T}A + \frac{128LT}{\mu^3 S_T}B. \tag{3.21}$$

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

3.6 Discussion

Qsparse-local-SGD asymptotically converges as fast as distributed vanilla SGD for $H = \mathcal{O}\left(\gamma T^{1/4}/(bR)^{3/4}\right)$ in the smooth and non-convex case and $H = \mathcal{O}\left(\gamma\sqrt{T/(bR)}\right)$ for the strongly convex case. Therefore this algorithm provides a lot of flexibility in terms of different ways of mitigating the communication bottleneck. For example, by increasing the batch size on each node, or by increasing the maximum synchronization period H up to allowable limits. Furthermore, one could also choose to opt for different values of k for the Top_k sparsifier, as well as adjust the parameter configurations of the quantizer, or in the most communication efficient case one could opt for the $SignTop_k$ composed operator. We present numerics in Chapter 5 demonstrating significant savings by a factor of 15-20 times over the state-of-the-art, in the number of bits exchanged. In our case, communication is of interest however one could present other operators satisfying the regularity condition, in Definition 7, and the analysis would still hold. For example in the extreme case, i.e., without compression, this generalizes to guarantees for local SGD.

CHAPTER 4

Distributed Asynchronous Operation

4.1 Asynchronous Operation

In a distributed setup, worker nodes may choose to update the master at arbitrary time intervals. It is also possible that the update schedule is decided by the master. In such a setup the worker nodes continuously perform updates on their local sequences and the master chooses a subset of workers to update from in each iteration. This is also useful when the data is generated at the worker nodes in a online manner which would require the master to synchronize with the corresponding workers only. We propose a general framework for such a setting where the workers are evolving sequences at the same rate and the extension to different processing speeds has been left as future work.

We propose and analyze a particular form of asynchronous operation where the workers synchronize with the master at arbitrary times decided locally or by master picking a subset of nodes as in federated learning [Kon17, MMR17]. 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. This is different from asynchronous algorithms studied for stragglers [WYL18, RRW11], where only one gradient step is taken but occurs at different times due to delays.

In this asynchronous setting, $\mathcal{I}_T^{(r)}$'s may be different for different workers. However, we assume that $gap(\mathcal{I}_T^{(r)}) \leq H$ holds for every $r \in [R]$, which means that there is a uniform bound on the maximum delay in each worker's update times. The algorithmic difference from

Algorithm 1 is that, in this case, *a subset of* workers (including a single worker) can send their updates to the master at their synchronization time steps; master aggregates them, updates the global parameter vector, and sends that only to those workers. Our algorithm is summarized in Algorithm 2

Algorithm 2 Qsparse-local-SGD with asynchronous updates

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
      $g_t^{(r)} \leftarrow Q\,Comp_k\left(m_t^{(r)} + \mathbf{x}_t^{(r)} - \widehat{\mathbf{x}}_{t+\frac{1}{2}}^{(r)}\right)$ 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

In this section we present our main convergence results with asynchronous updates, obtained by running the Algorithm 2 for smooth objectives, both non-convex and strongly

convex. All our results are under the smoothness assumption.

4.2 Main Results

Under the same assumption as in the synchronous setting, which are provided in Section 3.2, the following theorems hold even if Algorithm 2 is run with an arbitrary contraction operator whose contraction coefficient is equal to γ.

Theorem 3 (Convergence in the smooth (non convex) case with fixed learning rate)**.** *Under the same conditions as in Theorem 1 with* $gap(\mathcal{I}_T^{(r)}) \leq H$ *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.*

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

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

$$\mathbb{E}\|\nabla f(\mathbf{z}_T)\|_2^2 \leq \mathcal{O}\left(\frac{J\sigma_{max}}{\sqrt{bRT}}\right) + \mathcal{O}\left(\frac{J^2bRG^2}{\sigma_{max}^2\gamma^2T}(H^2 + H^4)\right), \tag{4.1}$$

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

Theorem 4 (Convergence in the smooth and strongly convex case with decaying learning rate)**.** *Under the same conditions as in Theorem 2 with* $gap(\mathcal{I}_T^{(r)}) \leq H$*, if* $\{\widehat{x}_t^{(r)}\}_{t=0}^{T-1}$ *is*

generated according to Algorithm 2, the following holds.

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \frac{La^3}{4S_T}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{8LT(T+2a)}{\mu^2 S_T}A + \frac{128LT}{\mu^3 S_T}D. \tag{4.2}$$

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

Corollary 5. *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* $\{\widehat{\boldsymbol{x}}_t^{(r)}\}_{t=0}^{T-1}$ *is generated according to Algorithm 2, the following holds:*

$$\mathbb{E}[f(\overline{\mathbf{x}}_T)] - f^* \leq \mathcal{O}\left(\frac{G^2H^3}{\mu^2\gamma^3T^3}\right) + \mathcal{O}\left(\frac{\sigma_{max}^2}{\mu^2bRT} + \frac{H\sigma_{max}^2}{\mu^2bR\gamma T^2}\right) + \mathcal{O}\left(\frac{G^2}{\mu^3\gamma^2T^2}(H^2+H^4)\right). \tag{4.3}$$

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

4.3 Proof Outline

Our proofs of Theorem 3 and Theorem 4 follow exactly along the lines of the proofs of Theorem 1 and Theorem 2 in the synchronous setting, but some technical details change significantly, which arise because, in our asynchronous setting, workers are allowed to update the global parameter vector in between two consecutive synchronization time steps of other workers.

In this asynchronous implementation, for a given worker r, $\mathcal{I}_T^{(r)} = \{t_{(i)}^{(r)} : i \in \mathbb{Z}^+, t_{(i)}^{(r)} \in [T], |t_{(i)}^{(r)} - t_{(j)}^{(r)}| \leq H\, \forall |i-j| \leq 1\}$. Following the same proof outlines as earlier, both (3.9) and (3.16) hold for smooth (non-convex) and strongly convex settings respectively. Now we provide the bounds for the deviation of local sequences in Lemma 9-10, as well as the difference between the virtual and true sequences in Lemma 11-12 for our asynchronous operation.

In Lemma 9-10, we show a bound on $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t-\widehat{\mathbf{x}}_t^{(r)}\|_2^2\leq\mathcal{O}(\eta_t^2G^2(H^2+H^4/\gamma^2)$, which is weaker than the corresponding bound $\mathcal{O}(\eta_t^2G^2H^2)$ for the synchronous setting. The proofs are provided in Appendix D.1 and D.2.

Lemma 9 (Contracting Local Sequence Deviation)**.** *For $\widehat{\boldsymbol{x}}_t,\widehat{\boldsymbol{x}}_t^{(r)}$ generated according to Algorithm 2 and gap$(\mathcal{I}_T^{(r)})\leq H$ the following holds*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\boldsymbol{x}}_t-\widehat{\boldsymbol{x}}_t^{(r)}\|_2^2\leq 8(1+C''H^2)\eta_t^2G^2H^2. \tag{4.4}$$

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

Lemma 10 (Bounded Local Sequence Deviation)**.** *For $\widehat{\boldsymbol{x}}_t,\widehat{\boldsymbol{x}}_t^{(r)}$generated according to Algorithm 2 with $\eta_t=\eta$ the following holds*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\boldsymbol{x}}_t-\widehat{\boldsymbol{x}}_t^{(r)}\|_2^2\leq(2+H^2C')\eta^2G^2H^2. \tag{4.5}$$

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

Now 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 $\mathbb{E}\|\widehat{\mathbf{x}}_t-\widetilde{\mathbf{x}}_t\|_2^2$. Note that in the synchronous case, we have shown in Lemma 6 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 $\mathbb{E}\|\widehat{\mathbf{x}}_t-\widetilde{\mathbf{x}}_t\|_2^2$ 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

$$\widehat{\mathbf{x}}_t-\widetilde{\mathbf{x}}_t=\left[\frac{1}{R}\sum_{r=1}^{R}\widehat{\mathbf{x}}_{t_r}^{(r)}-\bar{\bar{\mathbf{x}}}_{t_0'}\right]+\left[\bar{\bar{\mathbf{x}}}_{t_0'}-\bar{\bar{\mathbf{x}}}_t\right]+\left[\bar{\bar{\mathbf{x}}}_t-\frac{1}{R}\sum_{r=1}^{R}\widetilde{\mathbf{x}}_{t_r}^{(r)}\right] \tag{4.6}$$

In (4.6), the third term on the RHS is equal to the average memory as shown in (D.25) in Appendix D.3. Therefore unlike the synchronous setting, Lemma 6, i.e., $\widehat{\mathbf{x}}_t-\widetilde{\mathbf{x}}_t=\frac{1}{R}\sum_{r=1}^{R}m_t^{(r)}$ does not hold here; however, we 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 $\mathbb{E}\|\widehat{\mathbf{x}}_t-\widetilde{\mathbf{x}}_t\|_2^2\leq\mathcal{O}\left(\eta_t^2/\gamma^2G^2(H^2+H^4)\right)$, proved in Lemma 11-12 in Appendix D.3 and D.4, in comparison to $\mathcal{O}\left(\eta_t^2/\gamma^2G^2H^2\right)$ for the synchronous setting.

Lemma 11 (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*

$$\mathbb{E}\|\widehat{\boldsymbol{x}}_t - \widetilde{\boldsymbol{x}}_t\|_2^2 \leq C'\eta_t^2 H^4 G^2 + 12C\frac{\eta_t^2}{\gamma^2}G^2H^2. \tag{4.7}$$

Here $C' = 192(4-2\gamma)\left(1 + \frac{C}{\gamma^2}\right)$.

Lemma 12 (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* $\eta_t = \eta$, $gap(\mathcal{I}_T^{(r)} \leq H)$ *and* $t \in \mathbb{Z}^+$ *we have*

$$\mathbb{E}\|\widehat{\boldsymbol{x}}_t - \widetilde{\boldsymbol{x}}_t\|_2^2 \leq 6C'\eta^2 H^4 G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2. \tag{4.8}$$

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

4.4 Discussion

In our model of asynchrony we capture the arbitrary synchronization schedule decided by each worker node. The nodes evolve iterates locally which are synchronized with a global clock and the master chooses a subset of nodes to synchronize with, every once in a while. This model is relevant to the federated optimization setting, and as we did for the synchronous operation, we characterize the allowable limits of local computation while ensuring convergence at the rate of distributed vanilla SGD.

CHAPTER 5

Communication Cost and Experiments

5.1 Communication Cost

Infrequent communication as proposed in Algorithms 1 and 2 and their analysis provided in Chapter 3 and 4 characterizes the reduction in the amount of exchanges required between workers. What remains to be seen, is the bit expenses when workers communicate with the server in a parameter server framework, or among themselves when training over peer to peer networks. For our composed operators we observe significant reduction in bit expenditure, not only in comparison with full precision SGD or compressed versions of the same such as QSGD [AGL17], and signSGD [BWA18], but in fact it consumes less bandwidth than sparse methods that select Top_k coordinates [AHJ18, SCJ18]. This is because of the composition of stochastic and deterministic quantizers with a highly aggressive sparsifier such as Top_k.

We focused on the communication cost of gradient updates rather than aggregate broadcast. This is because the "broadcast" of updated parameters can be virtual in ring architectures (used in our numerics), where each node maintains the latest iterates internally, and exchanges the compressed gradient updates with other workers. The broadcast can also be inexpensive if the parameter server aggregates the sparse quantized updates and broadcasts it. Yet another scenario is one where the broadcast routine is implemented in a tree-structured manner (such as in many MPI implementations), relaxing the network bottleneck at the parameter server.

Besides the 32 bit scalar overhead required for the quantizer, the only expenses incurred are corresponding to the locations of the support of the updates and the number of levels of

quantization. We discuss the expenses corresponding to different quantizers and sparsifiers. The number of bits that are sent by a worker node to the master in a round of communication of the proposed algorithms when a composed operator QTop_k as in Lemma 1 is in use are:

1. *QSGD*: $(\lceil \log_2(d) \rceil + \lceil \log_2(s+1) \rceil + 1) \cdot k + 32$ which can be as low as $(\lceil \log_2(d) \rceil + \lceil \log_2(\sqrt{k}+1) \rceil + 1) \cdot k + 32$. For simplicity this is stated without the recursive Elias coding, as it is sufficient for us to comment on the gains via quantization [AGL17].

2. *Stochastic k-level Quantization*: $(\lceil \log_2(d) \rceil + \lceil \log_2(s+1) \rceil) \cdot k + 32$ which can be as low as $(\lceil \log_2(d) \rceil + \lceil \log_2(\sqrt{\frac{k}{2}}+1) \rceil) \cdot k + 32$

3. *Stochastic Rotated Quantization:* $(\lceil \log_2(d) \rceil + \lceil \log_2(s+1) \rceil) \cdot k + 32$ which can be as low as $(\lceil \log_2(d) \rceil + \lceil \log_2(\sqrt{2\log_2(2k)}+1) \rceil) \cdot k + 32$.

4. *No quantizer:* $(\lceil \log_2(d) \rceil + 32) \cdot k$.

For the scaled operator ${}^{\mathrm{QTop}_k}/_{1+\beta_{k,s}}$ as in Lemma 2, there are no restrictions on the operating regime in fact this also works for a very coarse quantizer (i.e., $\beta_{k,s} \gg 1$), though the convergence could be slow numerically. This is because the scaling factor $1+\beta_{k,s}$ is $\mathcal{O}(p(k))$ where $p(\cdot)$ is polynomial or log polynomial usually. Maximum savings are observed for the sign based operator in Lemma 3 when composed with $\mathrm{Comp}_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$. That is per round of communication the expenses incurred per worker is $(\lceil \log_2(d) \rceil + 1) \cdot k + 32$ bits.

For the Rand_k sparsifier, the parameter server can decide the random seeds for each worker and communicate them to each worker right after synchronization. While this operator is not as effective as Top_k, it asymptotically matches the rates of Top_kSGD, while at the same time affords less communication as the support locations need not be encoded with the updates. Therefore for the $SignRand_k$ operator, the expenses would be reduced to $k+32$ bits.

Note that the counting provided here is for when a single contraction operator is applied on the entire update. However from Corollary 1 we know that this would have to be more granular when different contraction operators are being used on different parts of the update.

5.2 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 5.1 and Table 5.2) 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 5.1 and Table 5.2).

		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 5.1 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 5.2 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.

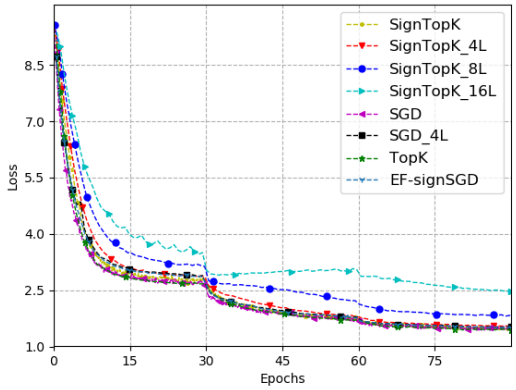
5.3 Non Convex Objective

5.3.1 Experiment setup

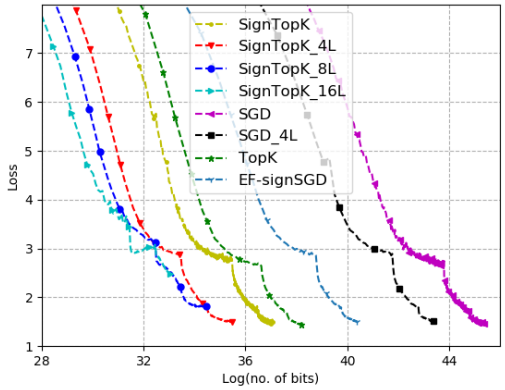
We train ResNet-50 [HZR16] (which has $d = 25,610,216$ parameters) on ImageNet dataset, using 8 NVIDIA Tesla V100 GPUs. We use a learning rate schedule consisting of 5 epochs of linear warmup, followed by a piecewise decay of 0.1 at epochs 30, 60 and 80, with a batch size of 256 per GPU. For the purposes of this thesis, we focus on SGD with momentum of 0.9, applied on the local iterations of the workers. We build our compression scheme into Horovod framework [SB18]. For quantization, we use $Sign$ operator as in Lemma 3. We use Top_k sparsification, and 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.

5.3.2 Results

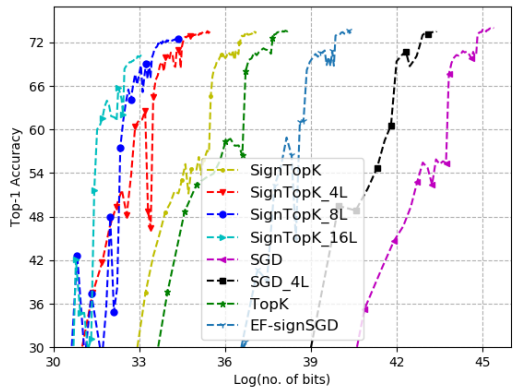
From Figure 5.1a, we observe that quantization and sparsification, when error compensation is enabled through accumulating errors, has almost no penalty in terms of convergence rate, with respect to vanilla SGD. Figure 5.1b, Figure 5.1c and Figure 5.1d show the training loss, top-1 and top-5 convergence rates respectively, with respect to the total number of bits of communication used. Here top-i refers to the accuracy of the top i predictions by the model from the list of possible classes, see [LHS15]. We observe that *Qsparse-local-SGD* combines the bit savings of the deterministic sign based operator and aggressive sparsifier along with infrequent communication, thereby outperforming the cases where these techniques are individually used. We exclude comparisons with stochastic quantizers such as in [AGL17, WXY17, SYK17],which are without any explicit sparsification, both for the non-convex and convex case, as their performance is much worse, see [SCJ18]. In particular, the required number of bits to achieve the same loss or top-1 accuracy in the case of *Qsparse-local-SGD* is around 1/16 in comparison with TopK-SGD and 1000× less than vanilla SGD. This also verifies that error compensation through memory can be used to mitigate not only



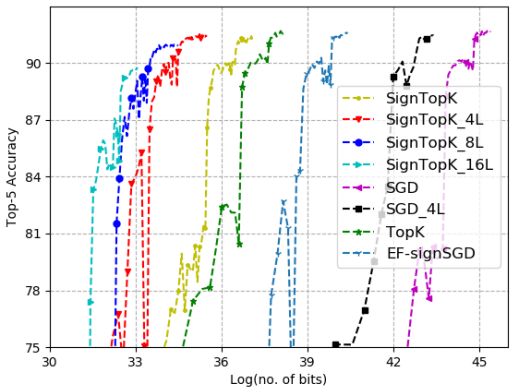
(a) Comparison of training loss against epochs



(b) Comparison of training loss with communication budget



(c) top-1 accuracy [LHS15] for schemes in Figure 5.1a



(d) top-5 accuracy [LHS15] for schemes in Figure 5.1a

Figure 5.1 Figures 5.1a-5.1c demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19], TopK-SGD [SCJ18, AHJ18] and local SGD [Sti19, YYZ18] in a non convex setting.

the missing components from updates in previous synchronization rounds, but also explicit quantization error.

5.4 Convex Objective

The experiments in figures 5.2a-5.2c are in a synchronous distributed setting with 15 workers each processing a mini-batch size of 8 samples per iteration using the *MNIST* [LBB98]

handwritten digits data set. The corresponding experiments for the asynchronous operation as in Algorithm 2 are shown in figures 5.3a-5.3b.

5.4.1 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 softmax regression with a standard ℓ_2 regularizer. The cost function is

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

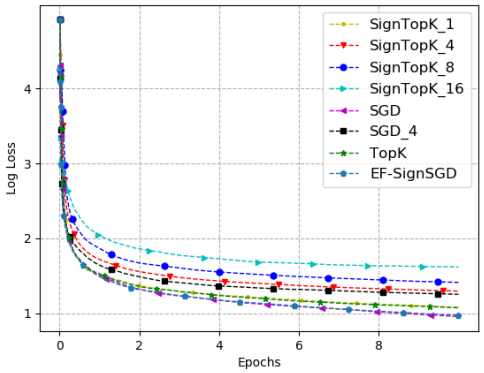
where $a^{(i)} \in \mathbb{R}^d$, $b^{(i)} \in [L]$ are the data points, which can belong to one of the L classes, and $\mathbf{x}_j \in \mathbb{R}^d$ for every $j \in [L]$, are columns of the parameter structured as follows

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

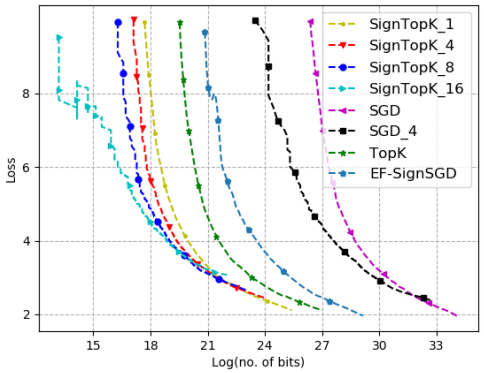
and $z^{(i)}$ for every $i \in [L]$ are the biases to be learnt corresponding to every class. We set λ to $1/n$.

5.4.2 Parameter selection and Learning rates

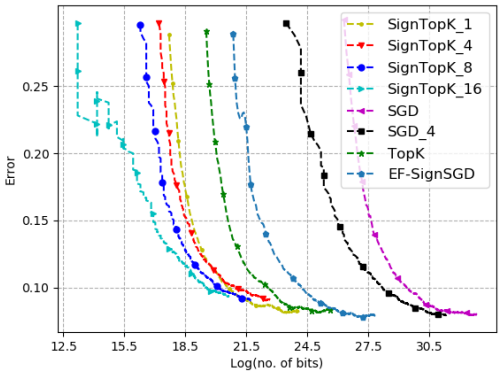
We use the deterministic operator as in Lemma 3 as our quantization method and Top_k with error compensation as the sparsifier. The schemes we compare with our composed $SignTop_k$ operator are EF-SIGNSGD [KRS19], TopK-SGD [SCJ18, AHJ18] and local SGD [Sti19]. The learning rate used for training is of the form $\frac{c}{\lambda(a+t)}$, where (i) λ is the regularization parameter; (ii) c is set with a careful hyperparameter sweep; (iii) $w_t = (a+t)^2$ as in Theorem 2, where a is set as $\frac{dH}{k}$ with d being the size of the gradient (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; (viii) $R = 15$ is the number of workers.



(a) Comparison of training loss against epochs between our scheme and other state of the art



(b) Comparison of training loss with the communication budget for schemes in Figure 5.2a



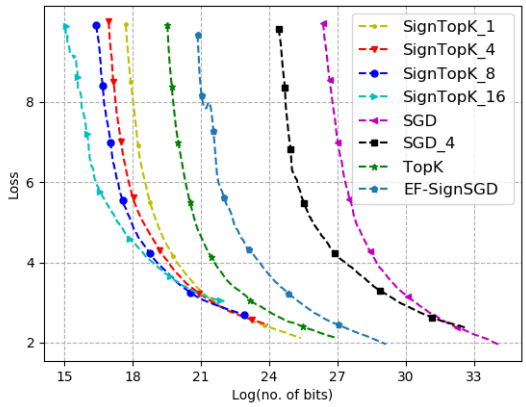
(c) Test Error using a model trained for given number of iterations, as seen in Figure 5.2a

Figure 5.2 Figures 5.2a-5.2c demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19] and TopK-SGD [SCJ18, AHJ18] in a convex setting for synchronous updates. Here for $H = 1, 4, 8, 16$, corresponds to the Algorithm 1 running with a synchronization period of at most H.

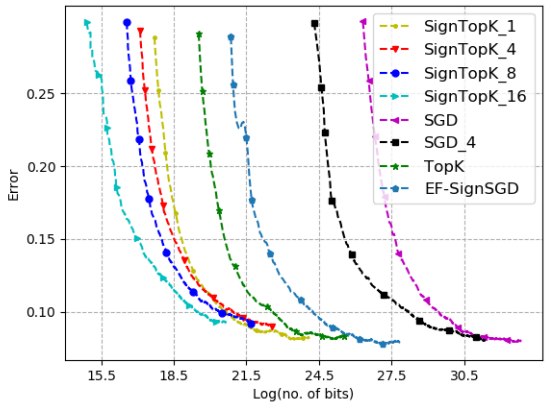
5.4.3 Experiment Results

In Figures 5.2a and 5.2b we compare the convergence of our proposed scheme in Algorithm 1, $SignTop_k$, with vanilla SGD (32 bit floating point), EF-SIGNSGD [KRS19] and TopK-SGD [SCJ18, AHJ18]. Both figures follow similar trends in which we observe $SignTop_k$ and TopK-SGD to be converging at the same rate as that of vanilla SGD, which is similar to the

observations in [SCJ18]. This implies that the composition of quantization with sparsification does not affect the convergence while achieving improved communication efficiency as can be seen in Figure 5.2c and 5.2b. Figure 5.2c shows that for test error approximately 0.1, *Qsparse-local-SGD* combines the benefits of the composed operator $SignTop_k$ with local computations and needs a factor of 10-15 times total bits less than TopK-SGD and 1000$\times$ less bits than vanilla SGD. We observe similar trends in Figures 5.3a-5.3b for our asynchronous operation where workers synchronize with the master at arbitrary time intervals as per Algorithm 2.



(a) Comparison of 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 5.3a

Figure 5.3 Figures 5.3a-5.3b demonstrate the performance of our scheme in comparison with EF-SIGNSGD [KRS19] and TopK-SGD [SCJ18, AHJ18] in a convex setting for asynchronous operation.

CHAPTER 6

Conclusion

In this thesis, we proposed a gradient compression scheme that composes both unbiased and biased quantization with aggressive sparsification. Furthermore we incorporated local computations which, when combined with quantization and explicit sparsification results in a highly communication efficient distributed algorithm, which we call *Qsparse-local-SGD*. We developed the convergence analyses of our scheme in both synchronous as well as asynchronous settings, and for both non-convex and convex objectives, and we showed that our proposed algorithm achieves the same rate as that of distributed vanilla SGD in each of these cases. Our schemes provide flexibility in terms of different options for mitigating the communication bottlenecks that arise in training high-dimensional learning models over bandwidth limited networks. When run without compression, this also subsumes/generalizes several recent results from the literature on local SGD, with similar convergence rates, as mentioned at the end of Section 3.4.

We believe that our approach for combining different forms of compression with local computations can easily be extended to the decentralized case, where nodes are communicate over an arbitrary connected graph, building on the ideas from [TGZ18,KRS19]. Our numerics also incorporate momentum acceleration, whose analysis is a topic for future research, for example, by incorporating ideas from [YJY19]. Although we use momentum for each local iteration, our preliminary results suggest that our method works with momentum applied to a block of updates as well though it was not the main focus of this thesis.

APPENDIX A

Supplementary material for preliminaries in Chapter 1

Proof of Remark 2

Proof. Take $g(\mathbf{x}) = f(\mathbf{x}) - \mathbf{x}^T\nabla f(\mathbf{y})$. $g : \mathbb{R}^d \to \mathbb{R}$ is convex with minima at $\mathbf{y}$. By L-smoothness, we can write

$$\begin{aligned} g(\mathbf{y}) &\leq g\left(\mathbf{x} - \frac{1}{L}\nabla g(\mathbf{x})\right) \\ &\leq g(\mathbf{x}) - \frac{1}{L}\|\nabla g(\mathbf{x})\|_2^2 + \frac{L}{2}\|\frac{1}{L}\nabla g(\mathbf{x})\|_2^2 \\ &= g(\mathbf{x}) - \frac{1}{2L}\|\nabla g(\mathbf{x})\|_2^2 \end{aligned} \tag{A.1}$$

Now substituting for $g(\mathbf{x})$ we get

$$f(\mathbf{y}) - \mathbf{y}^T\nabla f(\mathbf{y}) \leq f(\mathbf{x}) - \mathbf{x}^T\nabla f(\mathbf{y}) - \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

Substituting $y = \mathbf{x}^*$ we arrive at

$$\|\nabla f(\mathbf{x})\|_2^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^*))$$

This completes the proof of Remark 2. □

Proof of Remark 3

Proof. Take $g(\mathbf{x}) = \mathbf{x}^T\nabla f(\mathbf{y}) - f(\mathbf{x})$. $g : \mathbb{R}^d \to \mathbb{R}$ is concave with maxima at $\mathbf{y}$. By μ-strong convexity, we can write

$$\begin{aligned} g(\mathbf{y}) &\geq g\left(\mathbf{x} - \frac{1}{\mu}\nabla g(\mathbf{x})\right) \\ &\geq g(\mathbf{x}) - \frac{1}{\mu}\|\nabla g(\mathbf{x})\|_2^2 + \frac{\mu}{2}\|\frac{1}{\mu}\nabla g(\mathbf{x})\|_2^2 \\ &= g(\mathbf{x}) - \frac{1}{2\mu}\|\nabla g(\mathbf{x})\|_2^2 \end{aligned} \tag{A.2}$$

Now substituting for $g(\mathbf{x})$ we get

$$\mathbf{y}^T \nabla f(\mathbf{y}) - f(\mathbf{y}) \geq \mathbf{x}^T \nabla f(\mathbf{y}) - f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

Substituting $y = \mathbf{x}^*$ we arrive at

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2\mu (f(\mathbf{x}) - f(\mathbf{x}^*))$$

This completes the proof of Remark 3. □

Proof of Remark 4

Proof. For n arbitrary vectors $\{\mathbf{u}_i\}_{i=1}^n, \mathbf{u}_i \in \mathbb{R}^d$,

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{u}_i \right\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{u}_i\|_2^2$$

This follows from convexity and Jensen's inequality. Therefore we have

$$\left\| \sum_{i=1}^n \mathbf{u}_i \right\|_2^2 \leq n \sum_{i=1}^n \|\mathbf{u}_i\|_2^2 \tag{A.3}$$

This completes the proof of Remark 4. □

Proof of Remark 5

Proof. Let $\mathbf{X}$ be a random vector that takes on the value $\mathbf{u}_i$ with probability $1/n$ for all $i \in [n]$. Since variance is bounded by second moment we have,

$$\mathbb{E}\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2^2 \leq \mathbb{E}\|\mathbf{X}\|_2^2 \tag{A.4}$$

Also, $\mathbb{E}[\mathbf{X}] = \overline{\mathbf{u}}$, therefore we have

$$\frac{1}{n} \sum_{i=1}^n \|\mathbf{u}_i - \overline{\mathbf{u}}\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{u}_i\|_2^2 \tag{A.5}$$

This completes the proof of Remark 5. □

Proof of Remark 6

Proof. For any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and $\gamma > 0$,

$$\|\gamma^{\frac{1}{2}}\mathbf{u} - \gamma^{-\frac{1}{2}}\mathbf{v}\|_2^2 \geq 0$$

$$\gamma\|\mathbf{u}\|_2^2 + \gamma^{-1}\|\mathbf{v}\|_2^2 \geq 2\langle \mathbf{u}, \mathbf{v}\rangle \tag{A.6}$$

This completes the proof of Remark 6. □

APPENDIX B

Supplementary material for Chapter 2

B.1 Proof of Lemma 1

Lemma (Restating Lemma 1)**.** *Let $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ be a quantizer with parameter s that satisfies Definition 5. 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 $\mathbf{x} \in \mathbb{R}^d$. If k, s are such that $\beta_{k,s} < 1$. then $Q_sComp_k : \mathbb{R}^d \to \mathbb{R}^d$ is a contraction operator with the contraction coefficient being equal to $\gamma = (1 - \beta_{k,s})\frac{k}{d}$, i.e., for every $\mathbf{x} \in \mathbb{R}^d$, we have*

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

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

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

$$\begin{aligned}
&\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_sComp_k(\mathbf{x})\|_2^2] \\
&= \mathbb{E}_{C,Q}[\|\mathbf{x}\|_2^2] + \mathbb{E}_{C,Q}[\|Q_sComp_k(\mathbf{x})\|_2^2] \\
&\qquad\qquad - 2\mathbb{E}_C[\langle \mathbf{x}, \mathbb{E}_Q[Q_sComp_k(\mathbf{x})]\rangle] \\
&= \|\mathbf{x}\|_2^2 + \mathbb{E}_{C,Q}[\|Q_sComp_k(\mathbf{x})\|_2^2] - 2\mathbb{E}_C[\langle \mathbf{x}, Comp_k(\mathbf{x})\rangle]
\end{aligned}$$

In the last equality, we used that $\mathbf{x}$ is constant with respect to the randomness of Q_s and $Comp_k$, and that $\mathbb{E}_Q[Q_sComp_k(\mathbf{x})] = Comp_k(\mathbf{x})$, which follows from (i) of Definition 5. Observe that, for any $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$, we have $\langle \mathbf{x}, Comp_k(\mathbf{x})\rangle = \|Comp_k(\mathbf{x})\|_2^2$.

Continuing from above, we get

$$\mathbb{E}_{C,Q}[\|\mathbf{x} - Q_s Comp_k(\mathbf{x})\|_2^2] = \|\mathbf{x}\|_2^2 - 2\mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] + \mathbb{E}_{C,Q}[\|Q_s Comp_k(\mathbf{x})\|_2^2] \tag{B.1}$$

Observe that for any $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$, $Comp_k(\mathbf{x})$ 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 $\mathbb{E}_Q[\|Q_s Comp_k(\mathbf{x})\|_2^2] \leq (1+\beta_{k,s})\|Comp_k(\mathbf{x})\|_2^2$; see (ii) of Definition 5. Putting this back in (B.3), we get

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

Using $\mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] \geq \frac{k}{d}\|\mathbf{x}\|_2^2$ (see Lemma 13) in (B.4) gives

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

This completes the proof of Lemma 1. □

B.2 Proof of Lemma 2

Lemma (Restating Lemma 2)**.** *Let $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$. Let $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ be a stochastic quantizer with parameter s that satisfies Definition 5. Let $Q_s Comp_k : \mathbb{R}^d \to \mathbb{R}^d$ be defined as $Q_s Comp_k(\mathbf{x}) := Q_s(Comp_k(\mathbf{x}))$ for every $\mathbf{x} \in \mathbb{R}^d$. Then $\frac{Q_s Comp_k(\mathbf{x})}{1+\beta_{k,s}}$ is a contraction operator with the contraction coefficient being equal to $\gamma = \frac{k}{d(1+\beta_{k,s})}$, i.e., for every $\boldsymbol{x} \in \mathbb{R}^d$*

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

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

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

In (a) we used $\mathbb{E}_Q[Q_s Comp_k(\mathbf{x})] = Comp_k(\mathbf{x})$, which follows from (i) of Definition 5. Observe that, for $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$, we have $\langle \mathbf{x}, Comp_k(\mathbf{x})\rangle = \|Comp_k(\mathbf{x})\|_2^2$. Continuing from above, we get

$$\begin{aligned}\mathbb{E}_{C,Q}[\|\mathbf{x} - \frac{Q_s Comp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &= \|\mathbf{x}\|_2^2 - \frac{2}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\ &\quad + \frac{1}{(1+\beta_{k,s})^2}\mathbb{E}_{C,Q}\left[\|Q_s Comp_k(\mathbf{x})\|_2^2\right] \qquad \text{(B.3)}\end{aligned}$$

Observe that for any $Comp_k \in \{\text{Top}_k, \text{Rand}_k\}$, $Comp_k(\mathbf{x})$ 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 $\mathbb{E}_Q[\|Q_s Comp_k(\mathbf{x})\|_2^2] \leq (1+\beta_{k,s})\|Comp_k(\mathbf{x})\|_2^2$; see (ii) of Definition 5. Putting this back in (B.3), we get

$$\begin{aligned}\mathbb{E}_{C,Q}[\|\mathbf{x} - \frac{Q_s Comp_k(\mathbf{x})}{(1+\beta_{k,s})}\|_2^2] &\leq \|\mathbf{x}\|_2^2 - \frac{2}{1+\beta_{k,s}}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\ &\quad + \frac{1}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \\ &= \|\mathbf{x}\|_2^2 - \frac{1}{(1+\beta_{k,s})}\mathbb{E}_C\left[\|Comp_k(\mathbf{x})\|_2^2\right] \qquad \text{(B.4)}\end{aligned}$$

Using $\mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] \geq \frac{k}{d}\|\mathbf{x}\|_2^2$ (see (B.6) in Lemma 13) in (B.4) gives

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

This completes the proof of Lemma 2.

□

B.3 Proof of Lemma 3

Lemma (Restating Lemma 3). *For $Comp_k \in \{\mathrm{Top}_k, \mathrm{Rand}_k\}$, $\frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k}$, for any $m \in \mathbb{Z}_+$ is a contraction operator with the contraction coefficient γ_m being equal to*

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

For proving Lemma 3 we first state and prove Lemma 13 below.

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

$$\mathbb{E}[\|Comp_k(\mathbf{x})\|_1^2] \geq \max\left\{\frac{k}{d}\|\mathbf{x}\|_2^2, \frac{k^2}{d^2}\|\mathbf{x}\|_1^2\right\} \tag{B.5}$$

$$\mathbb{E}[\|Comp_k(\mathbf{x})\|_2^2] \geq \frac{k}{d}\|\mathbf{x}\|_2^2. \tag{B.6}$$

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

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

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

$$\mathbb{E}[\|\mathrm{Rand}_k(\mathbf{x})\|_1^2] \geq (\mathbb{E}[\|\mathrm{Rand}_k(\mathbf{x})\|_1])^2$$

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

□

Proof.

$$
\begin{aligned}
&\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \left\langle \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k}, \mathbf{x} \right\rangle + \|\mathbf{x}\|_2^2 \right] \\
&= \mathbb{E}_C \left[\frac{\|Comp_k(\mathbf{x})\|_m^2}{k} - 2 \frac{\|Comp_k(\mathbf{x})\|_m \|Comp_k(\mathbf{x})\|_1}{k} + \|\mathbf{x}\|_2^2 \right] \\
&\leq \|\mathbf{x}\|_2^2 - \frac{\mathbb{E}_C \|Comp_k(\mathbf{x})\|_m^2}{k} \qquad \text{(B.7)}
\end{aligned}
$$

In (B.7) 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 (B.5)) in (B.7) 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 (B.7) with $q = m$ and $p = 2$ gives

$$\begin{aligned}
&\mathbb{E}_C \left\| \frac{\|Comp_k(\mathbf{x})\|_m \, SignComp_k(\mathbf{x})}{k} - \mathbf{x} \right\|_2^2 \\
&\qquad \leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} \mathbb{E}_C[\|Comp_k(\mathbf{x})\|_2^2] \\
&\qquad \leq \|\mathbf{x}\|_2^2 - \frac{1}{k} k^{\frac{2}{m}-1} (k/d) \|\mathbf{x}\|_2^2 \quad \text{(By Lemma 13)} \\
&\qquad = \left[1 - \frac{k^{\frac{2}{m}-1}}{d} \right] \|\mathbf{x}\|_2^2. \qquad \text{(B.8)}
\end{aligned}$$

This completes the proof of Lemma 3. □

APPENDIX C

Supplementary material for Chapter 3

C.1 Proof of Lemma 4

Lemma (Restating Lemma 4)**.** *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 $a > \frac{4H}{\gamma}$, $\eta_t = \frac{\xi}{a+t}$, $gap(\mathcal{I}_T) \leq H$ and $t \in \mathbb{Z}^+$, there exists a $C \geq \frac{4a\gamma(1-\gamma^2)}{a\gamma-4H}$ such that*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta_t^2}{\gamma^2}CH^2G^2. \tag{C.1}$$

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 $\mathcal{I}_T = \{t_{(1)}, t_{(2)}, \ldots, t_{(l)} = T\}$. Fix any $i = 1, 2, \ldots, l$ 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 $r \in [R]$.

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

Here (a) is due to the contraction 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 $\|\mathbf{F}a+\mathbf{F}b\|^2 \leq (1+\tau)\|\mathbf{F}a\|^2+(1+\frac{1}{\tau})\|\mathbf{F}b\|^2$, which holds for every $\tau > 0$, in (C.2) gives (take any $p > 1$ in the following):

$$\begin{aligned}
\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2 &\leq (1-\gamma)\left[\left(1+\tfrac{(p-1)\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \left(1+\tfrac{p}{(p-1)\gamma}\right)\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2\right] \\
&\leq \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{(1-\gamma)(p\gamma+p)}{(p-1)\gamma}\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&= \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\mathbb{E}\|\widehat{\mathbf{x}}_{t_{(i)}}^{(r)} - \widehat{\mathbf{x}}_{t_{(i+1)}-\frac{1}{2}}^{(r)}\|^2 \\
&= \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\mathbb{E}\|\sum_{j=t_{(i)}}^{t_{(i+1)}-1}\eta_j\nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\leq \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\eta_{t_{(i)}}^2 H^2G^2 \qquad \text{(C.3)}
\end{aligned}$$

In the last inequality (C.3) 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:

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

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 $A = \xi^2H^2G^2$. Using this in (C.3) gives

$$\mathbb{E}\|m_{t_{(i+1)}}^{(r)}\|^2 \leq \left(1-\tfrac{\gamma}{p}\right)\mathbb{E}\|m_{t_{(i)}}^{(r)}\|^2 + \tfrac{p(1-\gamma^2)}{(p-1)\gamma}\tilde{\eta}_{t_{(i)}}^2 A. \qquad \text{(C.4)}$$

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 $m^{(r)}_{t_{(1)}-1} = m^{(r)}_0 = \mathbf{0}$. Consider the following:

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

Here (a) holds since $\mathbf{x}_{t_{(1)}-1} = \mathbf{x}_0 = \widehat{\mathbf{x}}^{(r)}_0$. It is easy to verify that $(1-\gamma)\tilde{\eta}_0^2 A \leq \frac{4a\gamma(1-\gamma^2)}{a\gamma - 4H}\frac{\tilde{\eta}^2_{t_{(1)}}}{\gamma^2}A$. 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 $\mathbb{E}\|m^{(r)}_{t_{(1)}}\|^2 \leq C\frac{\tilde{\eta}^2_{t_{(1)}}}{\gamma^2}A$.

Inductive case: Assume $\mathbb{E}\|m^{(r)}_{(i)}\|^2 \leq C\frac{\tilde{\eta}^2_{t_{(i)}}}{\gamma^2}A$ for some $i \in \mathbb{Z}^+$. We need to show that $\mathbb{E}\|m^{(r)}_{(i+1)}\|^2 \leq C\frac{\tilde{\eta}^2_{t_{(i+1)}}}{\gamma^2}A$. Using the inductive hypothesis in (C.4), we get

$$\begin{aligned}
\mathbb{E}\|m^{(r)}_{(i+1)}\|^2 &\leq \left(1 - \tfrac{\gamma}{p}\right) C\frac{\tilde{\eta}^2_{t_{(i)}}}{\gamma^2}A + \frac{p(1-\gamma^2)}{(p-1)\gamma}\tilde{\eta}^2_{t_{(i)}}A \\
&= C\frac{\tilde{\eta}^2_{t_{(i)}}}{\gamma^2}A\left(1 - \tfrac{\gamma}{p} + \frac{p(1-\gamma^2)}{p-1}\frac{\gamma}{C}\right) \\
&= C\frac{\tilde{\eta}^2_{t_{(i)}}}{\gamma^2}A\left(1 - \tfrac{\gamma}{p}\left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right)\right) \qquad \text{(C.5)}
\end{aligned}$$

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 $\tilde{\eta}^2_{t_{(i)}}\left(1 - \frac{\gamma}{p}\left(1 - \frac{p^2(1-\gamma^2)}{(p-1)C}\right)\right) \leq \tilde{\eta}^2_{t_{(i+1)}}$ 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 $\tilde{\eta}^2_{t_{(i)}+H} \leq \tilde{\eta}^2_{t_{(i+1)}}$), it suffices to show that $\tilde{\eta}^2_{t_{(i)}}\left(1 - \frac{\beta}{a}\right) \leq \tilde{\eta}^2_{t_{(i)}+H}$ holds whenever $\beta \geq 2H$. For simplicity of notation, let $t = t_{(i)}$. Note that $\tilde{\eta}^2_t\left(1 - \frac{\beta}{a}\right) = \frac{(a-\beta)}{a(a+t)^2}$. 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 (C.5) 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. □

C.2 Proof of Lemma 5

Lemma (Restating Lemma 5)**.** *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*

$$\mathbb{E}\|m_t^{(r)}\|^2 \leq 4\frac{\eta^2(1-\gamma^2)}{\gamma^2}H^2G^2 \tag{C.6}$$

Observe that (C.3) holds irrespective of the learning rate schedule. In particular, using

a fixed learning rate $\eta_t = \eta$ for every t gives

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

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

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

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

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

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

C.3 Proof of Lemma 6

Lemma (Restating Lemma 6)**.** *The memory is maintained so as to capture the distance between the true sequence and virtual sequence.*

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

Proof. 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 $t_r' + 1 \leq t_r$

$$\begin{aligned}\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\mathbf{x}}_{t_r+1}^{(r)} - \widetilde{\mathbf{x}}_{t_r+1}^{(r)}\right) \\ &= \frac{1}{R}\sum_{r=1}^{R}\left(\mathbf{x}_{t_r} - \frac{1}{R}\sum_{r=1}^{R} g_{t_r}^{(r)} - (\widetilde{\mathbf{x}}_{t_r'+1}^{(r)} - (\widehat{\mathbf{x}}_{t_r'+1}^{(r)} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}^{(r)}))\right)\end{aligned} \tag{C.8}$$

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

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^{R}\left(\mathbf{x}_{t_r} - \frac{1}{R}\sum_{r=1}^{R} g^{(r)}_{t_r} - (\widetilde{\mathbf{x}}^{(r)}_{t'_r+1} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}^{(r)}_{t_r+\frac{1}{2}}))\right) \\
&= \mathbf{x}_{t'_r+1} - \frac{1}{R}\sum_{r=1}^{R} g^{(r)}_{t_r} - (\widetilde{\mathbf{x}}_{t'_r+1} - (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}})) \\
&= \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t'_r+1} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}) - \frac{1}{R}\sum_{r=1}^{R} g^{(r)}_{t_r} \qquad \text{(C.9)}
\end{aligned}$$

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

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \widehat{\mathbf{x}}_{t'_r+1} - \widetilde{\mathbf{x}}_{t'_r+1} + (\mathbf{x}_{t_r} - \widehat{\mathbf{x}}_{t_r+\frac{1}{2}}) - \frac{1}{R}\sum_{r=1}^{R} g^{(r)}_{t_r} \qquad \text{(C.10)}$$

On rolling out the expression in (C.10) we get

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t &= \frac{1}{R}\sum_{r=1}^{R}\left[\sum_{\substack{j:j+1\in\mathcal{I}_T \\ j\le t_r}}\left(\mathbf{x}^{(r)}_j - \widehat{\mathbf{x}}^{(r)}_{j+\frac{1}{2}} - g^{(r)}_j\right)\right] \\
&= \frac{1}{R}\sum_{r=1}^{R} m^{(r)}_{t_r+1} \\
&= \frac{1}{R}\sum_{r=1}^{R} m^{(r)}_t \qquad \text{(C.11)}
\end{aligned}$$

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

C.4 Proof of Lemma 7

Lemma (Restating Lemma 7)**.** *With $\eta_t = \eta$ this follows from the analysis of Lemma 8*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}^{(r)}_t\|^2 \le \eta^2 G^2 H^2 \qquad \text{(C.12)}$$

Proof. Similar to analysis in (C.14) we can show that $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}^{(r)}_t\|^2 \le \eta^2 G^2 H^2$. □

C.5 Proof of Lemma 8

Lemma (Restating Lemma 8)**.** *Similar to Lemma 3.3 in [Sti19] we bound the deviation of the local sequences.*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2 \tag{C.13}$$

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.

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 &= \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)} - (\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_{t_r}^{(r)})\|^2 \\
&\leq \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \\
&\leq \eta_{t_r}^2 G^2 H^2 \\
&\leq 4\eta_t^2 G^2 H^2 \qquad \text{(C.14)}
\end{aligned}$$

The last inequality (C.14) uses $\eta_{t_r} \leq 2\eta_{t_r+H} \leq 2\eta_t$ and $t - t_r \leq H$. □

C.6 Smooth Objective: Proof of Theorem 1

Proof. Let $\mathbf{x}^*$ be the minimizer of $f(\mathbf{x})$, therefore we denote $f(\mathbf{x}^*)$ by f^*. For the purpose of reusing the proof later while proving Theorem 5, we start off with the decaying learning rate η_t until (C.18) and then switch to the fixed learning rate η. Note that the proof remains the same until (C.18) irrespective of the learning rate schedule; in particular, we can take $\eta_t = \eta$ and the same proof holds until (C.18).

By the definition of L-smoothness, we have

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

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

$$\begin{aligned}
\mathbb{E}_{i_t}[f(\widetilde{\mathbf{x}}_{t+1})] - f(\widetilde{\mathbf{x}}_t) &\leq -\frac{\eta_t}{2}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 + \|\frac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 - \|\nabla f(\widetilde{\mathbf{x}}_t) - \frac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2\right) \\
&\qquad + \eta_t^2 L\|\frac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 + \frac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 \\
&\leq -\frac{\eta_t}{2R}\sum_{r=1}^{R}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 - L^2\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2\right) + \frac{2\eta_t^2 L - \eta_t}{2}\|\frac{1}{R}\sum_{r=1}^{R}\nabla f^{(r)}(\widehat{\mathbf{x}}_t^{(r)})\|^2 \\
&\qquad + \frac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 \\
&= -\frac{\eta_t}{2R}\sum_{r=1}^{R}\left(\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 + L^2\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2\right) + \frac{2\eta_t^2 L - \eta_t}{2R}\sum_{r=1}^{R}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \\
&\qquad + \frac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + \frac{\eta_t L^2}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2. \qquad \text{(C.15)}
\end{aligned}$$

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

$$\begin{aligned}
\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 &\leq 2\|\nabla f(\widehat{\mathbf{x}}_t^{(r)}) - \nabla f(\widetilde{\mathbf{x}}_t)\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2 \\
&\leq 2L^2\|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t)\|^2, \qquad \text{(C.16)}
\end{aligned}$$

where the 2nd inequality follows from the smoothness (L-Lipschitz gradient) assumption.

Using this and that $\eta_t \leq \frac{1}{2L}$ in (C.15) and rearranging terms give

$$\frac{\eta_t}{4R}\sum_{r=1}^{R}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq f(\widetilde{\mathbf{x}}_t) - \mathbb{E}_{(i_t)}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + \frac{\eta_t L^2}{R}\sum_{r=1}^{R}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \quad \text{(C.17)}$$

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

$$\frac{\eta_t}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta_t^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 2\eta_t L^2\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 + 2\eta_t L^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \quad \text{(C.18)}$$

Observe that (C.18) holds irrespective of the learning rate schedule. In particular, if we take a fixed learning rate $\eta_t = \eta \leq \frac{1}{2L}$ in (C.18), we get

$$\frac{\eta}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 2\eta L^2\mathbb{E}\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 + 2\eta L^2\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \quad \text{(C.19)}$$

Lemma 6 and Lemma 5 together imply $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \frac{4\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2$. We also have from Lemma 7 that $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \eta^2 G^2 H^2$. Substituting these in (C.19) gives

$$\frac{\eta}{4R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \mathbb{E}[f(\widetilde{\mathbf{x}}_t)] - \mathbb{E}[f(\widetilde{\mathbf{x}}_{t+1})] + \frac{\eta^2 L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 8\frac{\eta^3(1-\gamma^2)}{\gamma^2}L^2G^2H^2 + 2\eta^3 L^2 G^2 H^2 \quad \text{(C.20)}$$

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

$$\frac{1}{4RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \frac{\mathbb{E}[f(\widetilde{\mathbf{x}}_0)] - f^*}{\eta T} + \frac{\eta L}{bR^2}\sum_{r=1}^{R}\sigma_r^2 + 8\frac{\eta^2(1-\gamma^2)}{\gamma^2}L^2G^2H^2 + 2\eta^2 L^2 G^2 H^2 \quad \text{(C.21)}$$

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

$$\frac{1}{RT}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^*}{\widehat{C}} + \frac{\widehat{C}L}{bR^2}\sum_{r=1}^{R}\sigma_r^2\right)\frac{4}{\sqrt{T}} + 8\left(4\frac{(1-\gamma^2)}{\gamma^2} + 1\right)\frac{\widehat{C}^2L^2G^2H^2}{T}. \quad \text{(C.22)}$$

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 $\Pr[\mathbf{z}_T = \widehat{\mathbf{x}}_t^{(r)}] = \frac{1}{RT}$, 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 (C.22) 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. □

C.7 Convex Objective: Proof of Theorem 2

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

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{\mathbf{x}}_t - \eta_t \mathbf{p}_t. \tag{C.23}$$

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 (C.23) we can get

$$\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 - 2\eta_t\left\langle\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t, \mathbf{p}_t - \overline{\mathbf{p}}_t\right\rangle \tag{C.24}$$

Taking the expectation w.r.t. the sampling i_t at time t (conditioning on the past) and noting that last term in (C.24) becomes zero gives:

$$\mathbb{E}_{i_t}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\mathbb{E}_{i_t}\|\mathbf{p}_t - \overline{\mathbf{p}}_t\|^2 \tag{C.25}$$

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

$$\mathbb{E}_{i_t}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 \leq \|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t\overline{\mathbf{p}}_t\|^2 + \eta_t^2\frac{\sum_{r=1}^{R}\sigma_r^2}{bR^2}. \tag{C.26}$$

Now we bound the first term on the RHS.

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

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 \leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{2L}(f(\widehat{\boldsymbol{x}}_t) - f^*) + \eta_t \left(\frac{3\mu}{2} + 3L\right) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \tag{C.27}$$

Proof.

$$\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 = \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 + \eta_t^2 \|\overline{\mathbf{p}}_t\|^2 - 2\eta_t \langle \widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t \rangle \tag{C.28}$$

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

$$\begin{aligned}
\|\overline{\mathbf{p}}_t\|^2 &= \|\frac{1}{R} \sum_{r=1}^{R} \left(\nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - \nabla f^{(r)}(\widetilde{\mathbf{x}}_t)\right) + \nabla f(\widetilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}^*)\|^2 \\
&\leq \frac{1}{R} \sum_{r=1}^{R} 2\|\nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - \nabla f^{(r)}(\widetilde{\mathbf{x}}_t)\|^2 + 2\|\nabla f(\widetilde{\mathbf{x}}_t) - \nabla f\left(\mathbf{x}^*\right)\|^2 \\
&\leq \frac{2L^2}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\| + 2\|\nabla f(\widetilde{\mathbf{x}}_t) - \nabla f\left(\mathbf{x}^*\right)\|^2
\end{aligned} \tag{C.29}$$

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 (C.29) gives

$$\eta_t^2 \|\overline{\mathbf{p}}_t\|^2 \leq \frac{2\eta_t^2 L^2}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t^{(r)} - \widetilde{\mathbf{x}}_t\| + 4\eta_t^2 L\left(f\left(\widetilde{\mathbf{x}}_t\right) - f(\mathbf{x}^*)\right) \tag{C.30}$$

Now we bound the last term of (C.28). By definition, we have

$$-2\eta_t \langle \widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t \rangle = -2\frac{\eta_t}{R} \sum_{r=1}^{R} \left\langle \widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \right\rangle - 2\frac{\eta_t}{R} \sum_{r=1}^{R} \left\langle \widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \right\rangle \tag{C.31}$$

For the first term on the RHS of (C.31), we can use strong convexity

$$-2\left\langle \widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \right\rangle \leq -2\left(f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - f^{(r)}(\mathbf{x}^*)\right) - \mu\|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \tag{C.32}$$

For the second term on the RHS of (C.31), we can use the following by smoothness.

$$-2\left\langle \widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}, \nabla f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) \right\rangle \leq L\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 + 2\left(f^{(r)}\left(\widehat{\mathbf{x}}_t^{(r)}\right) - f^{(r)}\left(\widetilde{\mathbf{x}}_t\right)\right) \tag{C.33}$$

Using (C.32)-(C.33) in (C.31) we get

$$
\begin{aligned}
-2\eta_t \left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t \right\rangle &\leq -\frac{2\eta_t}{R} \sum_{r=1}^{R} \left(f^{(r)}\left(\widetilde{\mathbf{x}}_t\right) - f^{(r)}(\mathbf{x}^*) \right) - \frac{\eta_t \mu}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \\
&\quad + \frac{L\eta_t}{R} \sum_{r=1}^{R} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \\
&= -2\eta_t \left(f\left(\widetilde{\mathbf{x}}_t\right) - f(\mathbf{x}^*) \right) - \frac{\eta_t \mu}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 + L\frac{\eta_t}{R} \sum_{r=1}^{R} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2
\end{aligned} \tag{C.34}
$$

Adding (C.30) and (C.34) and using $a \geq 32L/\mu$ which implies $\eta_t \leq 1/4L$ yields

$$
\begin{aligned}
\eta_t^2 \|\overline{\mathbf{p}}_t\|^2 - 2\eta_t \left\langle \widetilde{\mathbf{x}}_t - \mathbf{x}^*, \overline{\mathbf{p}}_t \right\rangle &\leq -2\eta_t (1 - 2\eta_t L) \left(f\left(\widetilde{\mathbf{x}}_t\right) - f^* \right) - \frac{\eta_t \mu}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t^{(r)} - \mathbf{x}^*\|^2 \\
&\quad + \frac{L\eta_t + 2\eta_t^2 L^2}{R} \sum_{r=1}^{R} \|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \\
&\leq -\eta_t \left(f\left(\widetilde{\mathbf{x}}_t\right) - f^* \right) - \eta_t \mu \|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \frac{3L\eta_t}{R} \sum_{r=1}^{R} \left(\|\widetilde{\mathbf{x}}_t - \widehat{\mathbf{x}}_t\|^2 + \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \right)
\end{aligned} \tag{C.35}
$$

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

$$
-\|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \leq \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{1}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \tag{C.36}
$$

Using (C.36) in (C.35) and then substituting (C.35) in (C.28) gives

$$
\begin{aligned}
\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \eta_t \left(f\left(\widetilde{\mathbf{x}}_t\right) - f^* \right) \\
&\quad + \eta_t (\mu + 3L) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2
\end{aligned} \tag{C.37}
$$

Using strong convexity of f we have

$$
\begin{aligned}
\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right) \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t \mu}{2} \|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\
&\quad + \eta_t (\mu + 3L) \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R} \sum_{r=1}^{R} \|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2
\end{aligned} \tag{C.38}
$$

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

$$\begin{aligned}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^* - \eta_t \overline{\mathbf{p}}_t\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t\mu}{4}\|\widehat{\mathbf{x}}_t - \mathbf{x}^*\|^2 \\ &\quad + \eta_t\left(\frac{3\mu}{2} + 3L\right)\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \\ &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t\mu}{2L}(f(\widehat{\mathbf{x}}_t) - f^*) \quad \text{(Using smoothness of } f(\mathbf{x})) \\ &\quad + \eta_t\left(\frac{3\mu}{2} + 3L\right)\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3L\eta_t}{R}\sum_{r=1}^{R}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \qquad \text{(C.39)}\end{aligned}$$

This completes the proof of Lemma 14. □

Using (C.39) in (C.26) and then taking the expectation over the entire process gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\eta_t\mu}{2L}(\mathbb{E}[f(\widehat{\mathbf{x}}_t)] - f^*) \\ &\quad + \eta_t\left(\frac{3\mu}{2} + 3L\right)\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 + \frac{3\eta_t L}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 + \eta_t^2\frac{\sum_{r=1}^{R}\sigma_r^2}{bR^2} \qquad \text{(C.40)}\end{aligned}$$

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 6 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 (C.40) and letting $e_t = \mathbb{E}[f(\widehat{\mathbf{x}}_t) - f^*]$ gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\mu\eta_t}{2L}e_t + \eta_t\left(\frac{3\mu}{2} + 3L\right)C\frac{4\eta_t^2}{\gamma^2}G^2H^2 \\ &\quad + (3L\eta_t)4\eta_t^2 LG^2H^2 + \eta_t^2\frac{\sum_{r=1}^{R}\sigma_r^2}{bR^2} \qquad \text{(C.41)}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}^*\|^2 &\leq \left(1 - \frac{\mu\eta_t}{2}\right)\mathbb{E}\|\widetilde{\mathbf{x}}_t - \mathbf{x}^*\|^2 - \frac{\mu\eta_t}{2L}e_t + \eta_t\left(\frac{3\mu}{2} + 3L\right)C\frac{4\eta_t^2}{\gamma^2}G^2H^2 \\ &\quad + (3\eta_t L)4\eta_t^2 LG^2H^2 + \eta_t^2\frac{\sum_{r=1}^{R}\sigma_r^2}{bR^2} \qquad \text{(C.42)}\end{aligned}$$

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

$$a_{t+1} \leq \left(1 - \frac{\mu\eta_t}{2}\right)a_t - \frac{\mu\eta_t}{2L}e_t + \eta_t^2 A + \eta_t^3 B \qquad \text{(C.43)}$$

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

$$\frac{\mu}{2LS_T} \sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu a^3}{8S_T} a_0 + \frac{4T\left(T+2a\right)}{\mu S_T} A + \frac{64T}{\mu^2 S_T} B \tag{C.44}$$

From convexity we can finally write

$$\mathbb{E} f\left(\overline{\mathbf{x}}_T\right) - f^* \leq \frac{La^3}{4S_T} a_0 + \frac{8LT\left(T+2a\right)}{\mu^2 S_T} A + \frac{128LT}{\mu^3 S_T} B \tag{C.45}$$

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

APPENDIX D

Supplementary material for Chapter 4

Maintain virtual sequences for every worker

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

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. $\mathcal{I}_T^{(r)} = \{t_{(i)}^{(r)} : i \in \mathbb{Z}^+, t_{(i)}^{(r)} \in [T],\ |t_{(i)}^{(r)} - t_{(j)}^{(r)}| \le H, \forall |i-j| \le 1\}$

D.1 Proof of Lemma 9

Lemma (Restating Lemma 9)**.** *For $\widehat{\boldsymbol{x}}_t, \widehat{\boldsymbol{x}}_t^{(r)}$ generated according to Algorithm 2 and $gap(\mathcal{I}_T^{(r)}) \le H$ the following holds*

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

Here $C'' = 8B(1 + \frac{C}{\gamma^2})$ where $B = 4 - 2\gamma$ and $C \ge \frac{4a\gamma(1-\gamma^2)}{a\gamma - 4H}$

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 $\sum_{i=1}^{n} \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 \leq \sum_{i=1}^{R} \|\mathbf{u}_i\|^2$. We use this in the first inequality below.

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

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

$$\begin{aligned}
\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 &= \mathbb{E}\|\sum_{j=t_r}^{t-1} \eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\leq (t-t_r)\sum_{j=t_r}^{t-1} \mathbb{E}\|\eta_j \nabla f_{i_j^{(r)}}\left(\widehat{\mathbf{x}}_j^{(r)}\right)\|^2 \\
&\leq (t-t_r)^2 \eta_{t_r}^2 G^2 \\
&\leq 4\eta_t^2 H^2 G^2. \qquad \text{(D.3)}
\end{aligned}$$

The last inequality (D.3) uses $\eta_{t_r} \leq 2\eta_{t_r+H} \leq 2\eta_t$ and $t - t_r \leq H$. To bound the second term of (D.2), note that we have

$$\bar{\bar{\mathbf{x}}}_{t_r}^{(r)} = \bar{\bar{\mathbf{x}}}_{t_0'} - \frac{1}{R}\sum_{s=1}^{R}\sum_{j=t_0'}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(s)}\} g_j^{(s)}. \qquad \text{(D.4)}$$

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 $\|\mathbb{1}\{j+1 \in \mathcal{I}_T^{(s)}\} g_j^{(s)}\|^2 \leq \|g_j^{(s)}\|^2$, we can upper-bound (D.4) as

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \leq \frac{(t_r - t_0')}{R}\sum_{s=1}^{R}\sum_{j=t_0'}^{t_r} \mathbb{E}\|g_j^{(s)}\|^2 \qquad \text{(D.5)}$$

Now we bound $\mathbb{E}\|g_j^{(s)}\|^2$ for any $j \in \{t_0', \ldots, t_r\}$ and $s \in [R]$: Since $\mathbb{E}\|QC(\mathbf{u})\|^2 \leq B\|\mathbf{u}\|^2$ holds for every $\mathbf{u}$, where $B = (4 - 2\gamma)$, which can be seen as follows: $\mathbb{E}\|QC(\mathbf{u})\|^2 \leq 2\mathbb{E}\|\mathbf{u} - QC(\mathbf{u})\|^2 + 2\|\mathbf{u}\|^2 \leq 2(1-\gamma)\|\mathbf{u}\|^2 + 2\|\mathbf{u}\|^2$, we have for any $s \in [R]$ that

$$\begin{aligned}
\mathbb{E}\|g_j^{(s)}\|^2 &\leq B\mathbb{E}\|m_j^{(s)} + \mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \qquad \text{(D.6)} \\
&\leq 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \qquad \text{(D.7)}
\end{aligned}$$

Observe that the proof of Lemma 4 does not depend on the synchrony of the network; it only uses the fact that $gap(\mathcal{I}_T^{(s)}) \leq H$ for any worker $s \in [R]$. Therefore, we can directly use Lemma 4 to bound the first term in (D.3) as $\mathbb{E}\|m_j^{(s)}\|^2 \leq 4C\frac{\eta_j^2}{\gamma^2}H^2G^2$. In order to bound the second term of (D.3), note that $\mathbf{x}_j^{(s)} = \widehat{\mathbf{x}}_{t_0'}^{(s)}$, which implies that $\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 = \|\sum_{l=t_0'}^{j}\eta_l \nabla f_{i_l^{(r)}}\left(\widehat{\mathbf{x}}_l^{(s)}\right)\|^2$. Taking expectation yields $\mathbb{E}\|\mathbf{x}_j^{(s)} - \widehat{\mathbf{x}}_{j+\frac{1}{2}}^{(s)}\|^2 \leq 4\eta_{t_0'}^2 H^2G^2$. Using these in (D.7) gives

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

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 (D.8)) in (D.5) gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\widehat{\mathbf{x}}}_{t_0'}\|^2 \leq 32B\left(1+\frac{C}{\gamma^2}\right)\eta_t^2 H^4G^2. \tag{D.9}$$

Putting this and the bound from (D.3) back in (D.2) gives

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

This completes the proof of Lemma 9. □

D.2 Proof of Lemma 10

Lemma (Restating Lemma 10)**.** *For $\widehat{\boldsymbol{x}}_t, \widehat{\boldsymbol{x}}_t^{(r)}$ generated according to Algorithm 2 with $\eta_t = \eta$ the following holds*

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\boldsymbol{x}}_t - \widehat{\boldsymbol{x}}_t^{(r)}\|^2 \leq (2+H^2C')\eta^2G^2H^2 \tag{D.10}$$

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

Proof. From (D.6) and (D.7) 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 $\mathbf{u}$

$$\begin{aligned}\mathbb{E}\|g_j^{(s)}\|^2 &\leq 2B\mathbb{E}\|m_j^{(s)}\|^2 + 2B\eta^2H^2G^2 \\ &\leq 8B\frac{(1-\gamma^2)\eta^2}{\gamma^2}H^2G^2 + 2\eta^2BH^2G^2 \\ &= 2B\left(\frac{4}{\gamma^2}-3\right)\eta^2H^2G^2 \end{aligned} \tag{D.11}$$

For a fixed learning rate η, using (D.11) and following similar analysis as in (D.3) we can bound the first term in (D.2) as follows

$$\mathbb{E}\|\widehat{\mathbf{x}}_t^{(r)} - \widehat{\mathbf{x}}_{t_r}^{(r)}\|^2 \leq \eta^2H^2G^2 \tag{D.12}$$

Similarly as in (D.4)-(D.8) we can bound the second term in (D.2) as follows

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \leq 2B\left(\frac{4}{\gamma^2}-3\right)\eta^2H^4G^2 \tag{D.13}$$

Using (D.12) and (D.13) in (D.2) we can show that

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq \left[2 + 4BH^2\left(\frac{4}{\gamma^2}-3\right)\right]\eta^2H^2G^2 \tag{D.14}$$

□

D.3 Proof of Lemma 11

Lemma (Restating Lemma 11)**.** *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*

$$\mathbb{E}\|\widehat{\boldsymbol{x}}_t - \widetilde{\boldsymbol{x}}_t\|^2 \leq C'\eta_t^2H^4G^2 + 12C\frac{\eta_t^2}{\gamma^2}G^2H^2 \tag{D.15}$$

Here $C' = 192B\left(1+\frac{C}{\gamma^2}\right)$ where $B = 4-2\gamma$.

Proof. Fix a time t and consider any worker $r \in [R]$. Let $t_r \in \mathcal{I}_T^{(r)}$ denote the last synchronization step until time t for the r'th worker. Define $t_0' := \min_{r\in[R]} t_r$. We want to

bound $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2$. Note that in the synchronous case, we have shown in Lemma 6 that $\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t = \frac{1}{R}\sum_{r=1}^R m_t^{(r)}$. 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

$$\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t = \left[\frac{1}{R}\sum_{r=1}^R \widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\right] + \left[\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\right] + \left[\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)}\right] \tag{D.16}$$

Applying Jensen's inequality and taking expectation gives

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \left[\frac{3}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2\right] + \left[3\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2\right] + \left[3\mathbb{E}\|\bar{\bar{\mathbf{x}}}_t - \frac{1}{R}\sum_{r=1}^R \widetilde{\mathbf{x}}_{t_r}^{(r)}\|^2\right] \tag{D.17}$$

We bound each of the three terms of (D.17) separately. We have upper-bounded the first term earlier in (D.9), which is

$$\mathbb{E}\|\widehat{\mathbf{x}}_{t_r}^{(r)} - \bar{\bar{\mathbf{x}}}_{t_0'}\|^2 \leq 32B\left(1 + \frac{C}{\gamma^2}\right)\eta_t^2 H^4 G^2. \tag{D.18}$$

To bound the second term of (D.17), note that

$$\bar{\bar{\mathbf{x}}}_t = \bar{\bar{\mathbf{x}}}_0 - \frac{1}{R}\sum_{r=1}^R \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \tag{D.19}$$

$$= \bar{\bar{\mathbf{x}}}_{t_0'} - \frac{1}{R}\sum_{r=1}^R \sum_{j=t_0'}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \tag{D.20}$$

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 (D.20) 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 (D.9) gives

$$\mathbb{E}\|\bar{\bar{\mathbf{x}}}_{t_0'} - \bar{\bar{\mathbf{x}}}_t\|^2 \leq 32B\left(1 + \frac{C}{\gamma^2}\right)\eta_t^2 H^4 G^2. \tag{D.21}$$

To bound the last term of (D.17), note that

$$\widetilde{\mathbf{x}}_{t_r}^{(r)} = \bar{\bar{\mathbf{x}}}_0 - \sum_{j=0}^{t_r-1} \eta_j \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right) \tag{D.22}$$

From (D.19) and (D.22), we can write

$$\bar{\bar{\mathbf{x}}}_t - \frac{1}{R} \sum_{r=1}^{R} \widetilde{\mathbf{x}}_{t_r}^{(r)} = \frac{1}{R} \sum_{r=1}^{R} \left[\sum_{j=0}^{t_r-1} \eta_j \nabla^{(r)} f_{(i_j)} \left(\widehat{\mathbf{x}}_j^{(r)} \right) - \sum_{j=0}^{t_r-1} \mathbb{1}\{j+1 \in \mathcal{I}_T^{(r)}\} g_j^{(r)} \right] \tag{D.23}$$

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

$$\mathbf{x}_{t_r^{(2)}-1}^{(r)} - \widehat{\mathbf{x}}_{t_r^{(2)}-\frac{1}{2}}^{(r)} = \sum_{j=t_r^{(1)}}^{t_r^{(2)}-1} \nabla f_{i_j^{(r)}} \left(\widehat{\mathbf{x}}_j^{(r)} \right). \tag{D.24}$$

Using (D.24) for every consecutive synchronization steps, we can equivalently write (D.23) as

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

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

$$\mathbb{E}\| \bar{\bar{\mathbf{x}}}_t - \frac{1}{R} \sum_{r=1}^{R} \widetilde{\mathbf{x}}_{t_r}^{(r)} \|^2 \le 4C \frac{\eta_t^2}{\gamma^2} G^2 H^2. \tag{D.26}$$

Putting the bounds from (D.18), (D.21), and (D.26) in (D.17) gives

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

This completes the proof of Lemma 11. □

D.4 Proof of Lemma 12

Lemma (Restating Lemma 12)**.** *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 $\eta_t = \eta$, $gap(\mathcal{I}_T^{(r)} \leq H)$ and $t \in \mathbb{Z}^+$ we have*

$$\mathbb{E}\|\widehat{\boldsymbol{x}}_t - \widetilde{\boldsymbol{x}}_t\|^2 \leq 6C'\eta^2 H^4 G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2 \tag{D.27}$$

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

Proof. For a constant learning rate the first term in (D.17) has been bounded earlier in (D.13). Following similar steps as in (D.20) we would have

$$\mathbb{E}\|\bar{\widehat{\mathbf{x}}}_{t'_0} - \bar{\widetilde{\mathbf{x}}}_t\|^2 \leq 2B\left(\frac{4}{\gamma^2}-3\right)\eta^2 H^4 G^2 \tag{D.28}$$

Finally using (D.13),(D.25), Lemma 5 and (D.28) in (D.17) we have

$$\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq 12B\left(\frac{4}{\gamma^2}-3\right)\eta^2 H^4 G^2 + \frac{12\eta^2(1-\gamma^2)}{\gamma^2}G^2H^2 \tag{D.29}$$

□

APPENDIX E

Supplementary material with additional results

Theorem 1 in Chapter 3 and Theorem 3 in Chapter 4 provide non asymptotic guarantees for Algorithm 1 and 2 respectively, where the horizon of training (T) is fixed, and the learning rate η is $\mathcal{O}(1/\sqrt{T})$. Here we provide results demonstrating asymptotic convergence with a decaying learning rate of $\xi/a+t$

E.1 Synchronous

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 contraction operator whose contraction coefficient is equal to $\gamma \in (0, 1]$. Let $\{\widehat{\boldsymbol{x}}_t^{(r)}\}_{t=0}^{T-1}$ 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.*

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

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.

E.2 Asynchronous

Theorem 6 (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 contraction operator*

whose contraction coefficient is equal to $\gamma \in (0,1]$. *Let* $\{\widehat{\boldsymbol{x}}_t^{(r)}\}_{t=0}^{T-1}$ *be generated according to Algorithm 2 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.*

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

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.

E.3 Proof of Theorem 5

Proof. Observe that we can use the proof of Theorem 1 exactly until (C.18), for $\eta_t \leq \frac{1}{2L}$ (which follows from our assumption that $a \geq 2\xi L$), which gives

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

We have from Lemma 8 that $\frac{1}{R}\sum_{r=1}^R \mathbb{E}\|\widehat{\mathbf{x}}_t - \widehat{\mathbf{x}}_t^{(r)}\|^2 \leq 4\eta_t^2 G^2 H^2$. Lemma 6 and Lemma 4 together imply that $\mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \leq \frac{1}{R}\sum_{r=1}^R \|m_t^{(r)}\|^2 \leq C\frac{4\eta_t^2}{\gamma^2}G^2H^2$. Using these bounds in (E.3) gives

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

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

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

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 (E.4) yields

$$\frac{1}{P_T}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\delta_t\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \leq \frac{\mathbb{E}f(\mathbf{x}_0)-f^*}{P_T} + \frac{L\xi^2}{bR^2(a-1)}\frac{\sum_{r=1}^{R}\sigma^2}{P_T} + \left(\frac{8C}{\gamma^2}+8\right)L^2G^2H^2\frac{\xi^3}{2P_T(a-1)^2} \tag{E.5}$$

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

$$\min_{t\in\{0,\ldots,T-1\},\, r\in[R]}\mathbb{E}\|\nabla f(\widehat{\mathbf{x}}_t^{(r)})\|^2 \xrightarrow{T\to\infty} 0. \tag{E.6}$$

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 (E.5)

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

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 $\sum_{t=0}^{T-1}\eta_t^3$:

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

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

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

This completes the proof of Theorem 5. □

[ABC16] 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 *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pp. 265–283, 2016.

[AGL17] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding." In *NIPS*, pp. 1707–1718, 2017.

[AH17] Alham Fikri Aji and Kenneth Heafield. "Sparse Communication for Distributed Gradient Descent." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 440–445, 2017.

[AHJ18] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli. "The Convergence of Sparsified Gradient Methods." In *NIPS*, pp. 5977–5987, 2018.

[BM11] Francis R. Bach and Eric Moulines. "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning." In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 451–459, 2011.

[Bot10] L. Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent." In *Proceedings of COMPSTAT'2010. Physica-Verlag HD*, 2010.

[BWA18] J. Bernstein, Y. Wang, K. Azizzadenesheli, and A. Anandkumar. "SignSGD: Compressed Optimisation for Non-Convex Problems." In *ICML*, pp. 559–568, 2018.

[CH16] Kai Chen and Qiang Huo. "Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering." In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 5880–5884. IEEE, 2016.

[Cop15] Gregory F. Coppola. *Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing.* PhD thesis, University of Edinburgh, UK, 2015.

[GMT73] R. Gitlin, J. Mazo, and M. Taylor. "On the design of gradient algorithms for digitally implemented adaptive filters." *IEEE Transactions on Circuit Theory*, **20**(2):125–136, March 1973.

[HK14] Elad Hazan and Satyen Kale. "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization." *Journal of Machine Learning Research*, **15**(1):2489–2512, 2014.

[HM51] Robbins Herbert and Sutton Monro. "A Stochastic Approximation Method." *The Annals of Mathematical Statistics. JSTOR, www.jstor.org/stable/2236626.*, **vol. 22, no. 3, pp. 400âĂŞ407**, 1951.

[HZR16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016.

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Kon17] Jakub Konecný. "Stochastic, Distributed and Federated Optimization for Machine Learning." *CoRR*, **abs/1707.01155**, 2017.

[KRS19] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. "Error Feedback Fixes SignSGD and other Gradient Compression Schemes." *CoRR*, **abs/1901.09847**, 2019.

[KSJ19] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. "Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication." *CoRR*, **abs/1902.00340**, 2019.

[LBB98] 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.

[LHM18] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training." In *ICLR*, 2018.

[LHS15] Maksim Lapin, Matthias Hein, and Bernt Schiele. "Top-k Multiclass SVM." In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 325–333, 2015.

[MMR17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data." In *AISTATS*, pp. 1273–1282, 2017.

[MPP17] 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.

[NJL09] 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.

[NND18] 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 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 3747–3755, 2018.

[RB93] M. Riedmiller and H. Braun. "A direct adaptive method for faster backpropagation learning: the RPROP algorithm." In *IEEE International Conference on Neural Networks*, pp. 586–591 vol.1, March 1993.

[RRW11] Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. "Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent." In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 693–701, 2011.

[RSS12] A. Rakhlin, O. Shamir, and K. Sridharan. "Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization." In *ICML*, 2012.

[SB18] A. Sergeev and M. D. Balso. "Horovod: fast and easy distributed deep learning in TensorFlow." *CoRR*, **abs/1802.05799**, 2018.

[SCJ18] S. U. Stich, J. B. Cordonnier, and M. Jaggi. "Sparsified SGD with Memory." In *NIPS*, pp. 4452–4463, 2018.

[SFD14] 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 2014*, pp. 1058–1062, 2014.

[SSS07] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM." In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pp. 807–814, 2007.

[Sti19] Sebastian U. Stich. "Local SGD Converges Fast and Communicates Little." In *ICLR*, 2019.

[Str15] Nikko Strom. "Scalable distributed DNN training using commodity GPU cloud computing." In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 1488–1492, 2015.

[SYK17] A. Theertha Suresh, F. X. Yu, S. Kumar, and H. B. McMahan. "Distributed Mean Estimation with Limited Communication." In *ICML*, pp. 3329–3337, 2017.

[TGZ18] Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. "Communication Compression for Decentralized Training." In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing*

Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada., pp. 7663–7673, 2018.

[TH12] T. Tieleman and G Hinton. *RMSprop. Coursera: Neural Networks for Machine Learning, Lecture 6.5.* 2012.

[WHH18] J. Wu, W. Huang, J. Huang, and T. Zhang. "Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization." In *ICML*, pp. 5321–5329, 2018.

[WJ18] Jianyu Wang and Gauri Joshi. "Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms." *CoRR*, **abs/1808.07576**, 2018.

[WSL18] H. Wang, S. Sievert, S. Liu, Z. B. Charles, D. S. Papailiopoulos, and S. Wright. "ATOMO: Communication-efficient Learning via Atomic Sparsification." In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 9872–9883, 2018.

[WWL18] J. Wangni, J. Wang, J. Liu, and T. Zhang. "Gradient Sparsification for Communication-Efficient Distributed Optimization." In *NIPS*, pp. 1306–1316, 2018.

[WXY17] 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*, pp. 1508–1518, 2017.

[WYL18] Tianyu Wu, Kun Yuan, Qing Ling, Wotao Yin, and Ali H. Sayed. "Decentralized Consensus Optimization With Asynchrony and Delays." *IEEE Trans. Signal and Information Processing over Networks*, **4**(2):293–307, 2018.

[YJY19] Hao Yu, Rong Jin, and Sen Yang. "On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization." In *Machine Learning, Proceedings of the Thirty-Sixth International Conference (ICML 2019)*, 2019.

[YYZ18] Hao Yu, Sen Yang, and Shenghuo Zhu. "Parallel Restarted SGD with Faster Convergence and Less Communication:Demystifying Why Model Averaging Works for Deep Learning." *CoRR*, **abs/1807.06629**, 2018.

[ZDJ13] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Information-theoretic lower bounds for distributed statistical estimation with communication constraints." In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pp. 2328–2336, 2013.

[ZDW13] Y. Zhang, J. C. Duchi, and M. J. Wainwright. "Communication-efficient algorithms for statistical optimization." *Journal of Machine Learning Research*, **14**(1):3321–3363, 2013.

[ZSM16] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. "Parallel SGD: When does averaging help?" *CoRR*, **abs/1606.07365**, 2016.