2020

Quantized SARAH

Siner Gokhan Yilmaz
Lehigh University

Quantized SARAH

by

Siner Gokhan Yilmaz

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Industrial and Systems Engineering

Lehigh University

May 2020

Thesis is accepted and approved in partial fulfillment of the requirements for the Master of Science in Industrial and System Engineering in Dept. of Industrial and Systems Engineering.

_____
Date

_____
Dr. Martin Takáč
Advisor

_____
Prof. Luis Nunes Vicente
Department Chair

Acknowledgments

I'm deeply grateful to my supervisor Dr. Martin Takáč for his significant contribution, guidance and great support throughout my study. I acquired valuable experience during my study under his supervision that will help me through my future career. I also want to thank my wife, Elif Irem Senyurt, for being there for me all the time.

Contents

List of Tables

List of Algorithms

List of Figures

Abstract

In this thesis, we propose the Quantized SARAH (QSARAH) algorithm, which is a modified version of SARAH, that can be used in the federated learning setting. A practical variant of QSARAH, QSARAH+, is also proposed, which uses diminishing step sizes property in the inner loop. The compression techniques used in our experiments are unbiased and have bounded second moment, which is crucial for QSARAH to work. The compression techniques we use enable us to achieve the same loss residual with a fewer number of bits communicated, which are supported by numerical experiments. The QSARAH has a linear convergence rate under convex and strongly convex assumptions.

Chapter 1

Introduction

1.1 Problem Definition

The emergence of interest in big data increased attention towards the empirical risk minimization problem. This problem can be formulated as

$$\min_{w \in \mathbb{R}^d} F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w), \tag{1.1}$$

where $w \in \mathbb{R}^d$ is the parameter vector, and we assume each $f_i,\ i \in \{1, \ldots, n\}$, is convex and lipschitz smooth. Also, we assume that there exists an optimal solution w^* to (1.1), which we do not have information beforehand.

The first intuitive approach would be applying traditional gradient descent (GD), whose update in each iteration is

$$w_{t+1} = w_t - \eta \nabla F(w_t),$$

where t is current iterate and $\nabla F(w_t)$ is full gradient, calculated at current iterate t. Since the function we are trying to optimize is smooth and convex, with carefully chosen η, GD produces iterates such that $F(w_t)$ converges to the optimal loss $F(w^*)$

at a linear rate. But GD is impractical in general since n is very large and it is very expensive to calculate the full gradient at each iteration.

As an alternative to GD, stochastic gradient descent[11] (SGD) is commonly used to solve (1.1). The update of SGD is similar to that of GD, which can be written as

$$w_{t+1} = w_t - \eta_t \nabla f_i(w_t),$$

where index i is chosen uniformly random from $\{1, \ldots, n\}$, that is, the update is made by using the gradient of one of the functions. Although its per iteration cost is n times better than GD, with constant η_t, the algorithm is guaranteed to converge only a neighborhood of w^*. This is caused by the variance of gradient estimate, although it is an unbiased estimate of full gradient. This effect can be diminished by choosing decreasing step sizes, but then, SGD can only achieve sublinear convergence with the strongly convex objective. Instead of using the gradient of a single function, one can get a randomly chosen set of functions to calculate the gradient estimate, which is called mini-batched SGD[15]. This method can reduce the variance of gradient estimate greatly, with a trade-off of increased computation cost per iteration.

There is another class of algorithms, which are called Variance Reduction methods. These methods utilize full gradient (or approximation of full gradient) snapshots in order to reduce the variance of current iterate. Their generalized update can be written in the form

$$w_{t+1} = w_t - \eta_t(\nabla f_i(w_t) - \xi_i),$$

where ξ_i is an estimation of $\nabla f_i(w_t) - \nabla f(w_t)$. The idea in this class of algorithms is to estimate the deviation of the stochastic gradient from the true gradient, and use it to get a better approximation of the full gradient. Some examples of this class of algorithms are SAG[12], SAGA[2], SVRG[6] and SARAH[9]. Our focus on this thesis

will be on a modified version of SARAH, but before going into more detail, we need to mention distributed learning and federated learning to further motivate our work.

1.2 Distributed Learning

With increasing amounts of data available, solving (1.1) on a single machine became infeasible for most of the applications, either because data cannot be stored on a single machine or amount of calculations needed cannot be achieved in a reasonable amount of time with a single machine. To overcome this problem, multiple nodes can be used to share computational resources and data storage, which is called distributed learning. Although there has been research on distributed learning for decades, the increasing model complexity and huge amounts of data made it very popular in the last decade. In distributed setting, communication costs are much higher compared to the cost of reading data from memory, which makes communication costs bottleneck in this setting[5].

1.3 Federated Learning

Technological improvements increased the computational power of personal devices such as mobile phones and tablets. These devices are also a valuable source of data, as they have rich in the sense of sensors. Even without sensors, user behavior can be used in a lot of applications, such as auto-completion of words. There is another setting introduced in order to utilize this additional computation power, which is called federated learning[7].

Like distributed learning, the main bottleneck in federated learning is communication costs, but it introduces additional challenges. These challenges are a result of privacy issues, ie. each device uses its data in its computations. Firstly, the number of nodes is much larger than the number of data points each sample has, ie. **Massively**

distributed. Secondly, since each device has different capabilities and the user is another variable factor, data has different distribution for each node, ie. **Non-IID**. Finally, each node can have a different number of data samples, ie. **Unbalanced**.

1.4 Communication Costs as Bottleneck

As we mentioned, in distributed learning and federated learning settings communication costs are the main bottleneck. The main idea to solve this problem is to increase the ratio between computation time and communication time. To increase this ratio, one approach is to increase computation that is made to calculate the gradient estimate, which results in a better convergence rate per communication round. One can use larger mini-batches[3], define harder sub-problems for nodes to solve[10] or reduce number of communication rounds per iteration[16] in order to achieve this.

Another approach in the literature is to reduce communicated bits while sacrificing accuracy. It is shown that this can be done by reducing the precision of the vectors being sent[13] or even sending signs of each coordinate can be enough[14]. Although these two methods are working, the bias introduced makes them hard to analyze. This problem can be overcome by quantizing vectors stochastically such that the expected value of update is unchanged[1, 4].

1.5 Contributions

In this thesis, we will introduce a variant of SARAH[9], which is suitable for the federated learning setting, namely Quantized SARAH (QSARAH). We will use different quantization methods when communicating updates and still prove convergence if those methods have certain stochastic properties. Also, we will show empirical results to demonstrate our method can have better convergence with less communication when using quantization methods.

Chapter 2

Quantized SARAH

2.1 Vanilla SARAH

Vanilla version of SARAH[9] consists of two loops, outer loop and inner loop. In the outer loop, full gradient is calculated and it is later used for one gradient descent step. In the inner loop, previous update and stochastic gradients for current and previous iterate are used to calculate new update, ie.

$$v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)},$$

which is followed by inner loop update, that is,

$$w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}.$$

Here, s is outer loop index, t is inner loop index and i_t is index for function that is chosen uniformly random from $\{0, 1, \ldots, m\}$ for each t. Overall algorithm can be seen at Algorithm 1.

Algorithm 1 SARAH

Parameters: The learning rate $\eta > 0$ and inner loop size m
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**
  $w_0^{(s)} = \tilde{w}_{s-1}$
  $v_0^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_0^{(s)})$
  $w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$
  Iterate:
  for $t = 1, 2, \ldots, m-1$ **do**
    Sample i_t uniformly at random from $\{1, 2, \ldots, n\}$
    $v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}$
    $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$
  end for
  Set $\tilde{w}_s = w_t^{(s)}$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
end for

2.2 Quantization

Before explaining Quantized SARAH, we will explain how quantization works and how this technique benefits us. Let us start with a more general definition, which is called *compression operators*.

Definition 2.1 (Compression operators, Definition 2 at [4])**.** A function $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ mapping a deterministic input to a random vector is called a *compression operator* (on $\mathbb{R}^d$). We say that $\mathcal{C}$ is unbiased if

$$\mathbb{E}[\mathcal{C}(w)] = w, \quad \forall w \in \mathbb{R}^d. \tag{2.1}$$

Moreover, if there exists $\omega \geq 0$ such that

$$\mathbb{E}[\|\mathcal{C}(w)\|^2] \leq (\omega + 1)\|w\|^2, \quad \forall w \in \mathbb{R}^d, \tag{2.2}$$

we say that $\mathcal{C}$ has bounded second moment. If $\mathcal{C}$ is unbiased and has bounded second moment, we will say $\mathcal{C} \in \mathbb{B}(\omega)$.

The quantization operations we use, *natural quantization* and *uniform quantization*[4], are special cases of *general dithering*. For $1 \leq p \leq +\infty$, let $\|w\|_p$ be p-norm: $\|w\|_p \coloneqq (\sum_i |w_i|^p)^{1/p}$.

Definition 2.2 (General dithering, Definition 3 at [4])**.** The *general dithering* operator with respect to the p norm and with s levels $0 = l_s < l_{s-1} < \cdots < l_1 < l_0 = 1$, denoted $\mathcal{D}^{\mathcal{C},p,s}_{gen}$, is defined as follows. Let $w \in \mathbb{R}^d$. If $w = 0$, we let $\mathcal{D}^{\mathcal{C},p,s}_{gen}(w) = 0$. If $w \neq 0$, we let $y_i \coloneqq |w_i|/\|w\|_p$ for all $i \in \{1, 2, \ldots, d\}$. Assuming $l_{u+1} \leq y_i \leq l_u$ for some $u \in \{0, 1, \ldots, s-1\}$, we let

$$(\mathcal{D}^{\mathcal{C},p,s}_{gen}(w))_i = \mathcal{C}(\|w\|_p) \times \operatorname{sign}(w_i) \times \xi_i(y_i), \tag{2.3}$$

where $\mathcal{C} \in \mathbb{B}(\omega)$ for some $\omega \geq 0$ and $\xi_i(y_i)$ is a random variable equal to l_u with probability $\frac{y_i - l_{u+1}}{l_u - l_{u+1}}$, and to l_{u+1} with probability $\frac{l_u - y_i}{l_u - l_{u+1}}$. Note that $\mathbb{E}[\xi_i(y_i)] = y_i$.

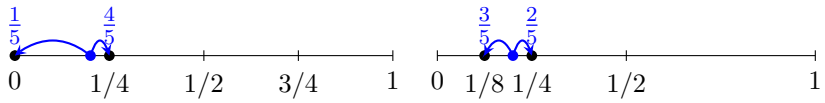

Figure 2.1: Random rounding example for uniform quantization (left) and natural quantization (right) for $y_i = \frac{1}{5}$ and $s = 4$.

Uniform quantization, $Q^s_{uni} : \mathbb{R}^d \to \mathbb{R}^d$, can be obtained by setting $l_{s-1} = 1/s, l_{s-2} = 2/s, \ldots, l_1 = (s-1)/s$, $p = 2$ and $\mathcal{C}$ equal to the identity operator. Similarly, natural quantization, $Q^s_{nat} : \mathbb{R}^d \to \mathbb{R}^d$, can be obtained by setting $l_{s-1} = 2^{1-s}, l_{s-2} = 2^{2-s}, \ldots, l_1 = 2^{-1}$, $p = 2$ and $\mathcal{C}$ equal to the identity operator. One example of random rounding (ξ_i) for uniform quantization and natural quantization can be seen on Figure 2.1.

By using these quantization techniques, instead of d floats, which is $32d$ bits, we are sending one float for the norm of the vector, $\lceil \log_2(s+1) \rceil d$ bits for quantized coordinates and d bits for signs of the coordinates. For example, for large d, quantization with $s = 7$ takes 8 times fewer bits to communicate. Quantization introduces some variance to vectors so there is a trade-off between quantization levels and convergence rate.

2.3 QSARAH

The outer iteration update in the QSARAH yields the same update as SARAH, the only difference is, it is done in parallel. Every node calculates the gradient of the data it has and sends it to update node, which later calculates the full gradient and sends back aggregated vector to the other nodes. Quantization in the outer loop does not make sense since the accuracy of this update is more important than inner loop updates, and time saved would not justify accuracy lost by quantization.

The inner iteration update, on the other hand, have a few differences. Instead of calculating gradients of current and previous iterates only at single data point, they are calculated on b different nodes. After that, each node quantizes the gradient difference and sends it to parameter server. Quantization operation is defined as $Q : \mathbb{R}^d \to \mathbb{R}^d$ and it must satisfy (2.1) and (2.2), in order for algorithm to converge. This quantization step is not only for decreasing communication costs, it is another layer of privacy measure for federated learning setting. Later in the parameter server, these quantized gradient differences are averaged and used for calculating overall gradient estimate. These steps can be written as,

$$v_t^{(s)} = \frac{1}{b} \sum_{i \in I_t} [Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))] + v_{t-1}^{(s)}, \tag{2.4}$$

which is later used for the iterate update

$$w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}. \tag{2.5}$$

Here, I_t is a set of indices, chosen uniformly at random from $\{1, 2, \ldots, n\}$. Overall algorithm can be seen at Algorithm 2.

Algorithm 2 Quantized SARAH

Parameters: The learning rate $\eta > 0$ and inner loop size m
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**
  $w_0^{(s)} = \tilde{w}_{s-1}$
  $v_0^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_0^{(s)})$
  $w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$
  Iterate:
  for $t = 1, 2, \ldots, m-1$ **do**
    Sample b samples, I_t, uniformly at random without replacement from $\{1, 2, \ldots, n\}$
    $v_t^{(s)} = \frac{1}{b} \sum_{i \in I_t} [\mathcal{C}_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))] + v_{t-1}^{(s)}$
    $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$
  end for
  Set $\tilde{w}_s = w_t^{(s)}$ with t chosen uniformly at random from $\{0, 1, \ldots, m\}$
end for

Similar to SARAH+[9], we also propose a practical variant for QSARAH, which is QSARAH+. Instead of using fixed m, we use similar condition as SARAH+ for terminating inner loop. Just like SARAH+, we use last iterate of inner loop, for the next outer loops starting point, instead of uniformly randomly choosing one of the inner iterates. Overall algorithm for QSARAH+ can be seen at Algorithm 3. We use QSARAH+ for our numerical experiments.

Algorithm 3 Quantized SARAH+

Parameters: The learning rate $\eta > 0$ and inner loop size m
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots$ **do**
$w_0^{(s)} = \tilde{w}_{s-1}$
$v_0^{(s)} = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_0^{(s)})$
$w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$
$t = 1$
while $\|v_{t-1}^{(s)}\|^2 > \gamma \|v_0^{(s)}\|^2$ **and** $t < m$ **do**
Sample b samples, I_t, uniformly at random without replacement from $\{1, 2, \ldots, n\}$
$v_t^{(s)} = \frac{1}{b} \sum_{i \in I_t} [\mathcal{C}_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))] + v_{t-1}^{(s)}$
$w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$
$t = t + 1$
end while
Set $\tilde{w}_s = w_t^{(s)}$
end for

Chapter 3

Theoretical Analysis

3.1 Quantization

In order to QSARAH to converge, quantization scheme we use should satisfy (2.1) and (2.2). Let us start with proving quantization techniques we use (Q^s_{uni} and Q^s_{nat}) have these properties.

Theorem 3.1. *$Q^s_{uni} \in \mathbb{B}(\omega)$ where $\omega = \min(d/(4s^2), \sqrt{d}/s)$.*

Proof. (2.1) holds since $\mathbb{E}[\xi_i(y_i)] = y_i$, $\forall i \in \{1, 2, \ldots, d\}$, from definition of general dithering. For (2.2), let us start with bounding $\mathbb{E}[\xi_i(y_i)^2]$,

$$\begin{aligned}
\mathbb{E}[\xi_i(y_i)^2] &= \mathbb{E}[\xi_i(y_i)]^2 + \mathbb{E}\big[(\xi_i(y_i) - \mathbb{E}[\xi_i(y_i)])^2\big] \\
&= y_i^2 + (l_u - y_i)^2 \frac{y_i - l_{u+1}}{l_u - l_{u+1}} + (y_i - l_{u+1})^2 \frac{l_u - y_i}{l_u - l_{u+1}} \\
&= y_i^2 + (l_u - y_i)(y_i - l_{u+1}),
\end{aligned}$$

note that $l_{u+1} \leq y_i \leq l_u$ and $l_u - l_{u+1} = 1/s$ and $0 \leq l_u, l_{u+1}, y_i \leq 1$. Since $\max_{y_i}((l_u - y_i)(y_i - l_{u+1})) = \frac{1}{4s^2}$ at $y_i = \frac{1}{2s} + l_u$ and $(l_u - y_i)(y_i - l_{u+1}) \leq \frac{1}{s}(y_i - l_{u+1}) \leq \frac{y_i}{s}$, we

can write

$$\begin{aligned}\mathbb{E}[\xi_i(y_i)^2] &\le y_i^2 + \min\Big(\frac{1}{4s^2}, \frac{y_i}{s}\Big) \\ &= \frac{w_i^2}{\|w\|^2} + \min\Big(\frac{1}{4s^2}, \frac{|w_i|}{s\|w\|}\Big).\end{aligned}$$

Using this result, we have

$$\begin{aligned}\mathbb{E}[\|Q^s_{uni}(w)\|^2] &= \sum_{i=1}^{d} \mathbb{E}\big[\|w\|^2 \xi_i(y_i)^2\big] \\ &\le \|w\|^2 \sum_{i=1}^{d} \left(\frac{w_i^2}{\|w\|^2} + \min\Big(\frac{1}{4s^2}, \frac{|w_i|}{s\|w\|}\Big)\right) \\ &\le \|w\|^2 \left(1 + \min\Big(\frac{d}{4s^2}, \frac{\|w\|_1}{s\|w\|}\Big)\right) \\ &\le \|w\|^2 \left(1 + \min\Big(\frac{d}{4s^2}, \frac{\sqrt{d}}{s}\Big)\right),\end{aligned}$$

where last inequality comes from $\max_w \frac{\|w\|_1}{\|w\|} = \sqrt{d}$. □

Theorem 3.2. *$Q^s_{nat} \in \mathbb{B}(\omega)$ where $\omega = \frac{1}{8} + 2^{1-s}\min(\sqrt{d}, 2^{1-s}d)$.*

Proof. (2.1) holds since $\mathbb{E}[\xi_i(y_i)] = y_i$, $\forall i \in \{1, 2, \ldots, d\}$, from definition of general dithering. For (2.2), let us start with bounding $\mathbb{E}[\xi_i(y_i)^2]$. First, for $y_i < 2^{1-s}$,

$$\begin{aligned}\mathbb{E}[\xi_i(y_i)^2] &= l_{s-1}^2 \frac{y_i - l_s}{l_{s-1} - l_s} + l_s^2 \frac{l_{s-1} - y_i}{l_{s-1} - l_s} \\ &= 2^{1-s} y_i,\end{aligned}$$

note that $l_s = 0$ and $l_u = 2^{-u}, u \in \{0, 1, \ldots, s-1\}$. For $y_i \ge 2^{1-s}$,

$$\begin{aligned}\mathbb{E}[\xi_i(y_i)^2] &= l_u^2 \frac{y_i - l_{u+1}}{l_u - l_{u+1}} + l_{u+1}^2 \frac{l_u - y_i}{l_u - l_{u+1}} \\ &= 2^{-2u} \frac{y_i - 2^{-u-1}}{2^{-u} - 2^{-u-1}} + 2^{-2u-2} \frac{2^{-u} - y_i}{2^{-u} - 2^{-u-1}} \\ &= 2^{1-u}(y_i - 2^{-u-1}) + 2^{-u-1}(2^{-u} - y_i)\end{aligned}$$

$$
\begin{aligned}
&= \frac{3y_i - 2^{-u}}{2} 2^{-u} \\
&\leq y_i^2 \max_{y_i} \frac{3y_i - 2^{-u}}{2y_i^2} 2^{-u} \\
&= \frac{9}{8} y_i^2,
\end{aligned}
$$

since $\max_{y_i} \frac{3y_i - 2^{-u}}{2y_i^2} 2^{-u} = \frac{9}{8}$ at $y_i = \frac{2^{1-u}}{3}$. Combining these two cases, for $0 \leq y_i \leq 1$

$$
\begin{aligned}
\mathbb{E}[\xi_i(y_i)^2] &\leq \mathbb{1}(y_i \geq 2^{1-s}) \frac{9}{8} y_i^2 + \mathbb{1}(y_i < 2^{1-s}) 2^{1-s} y_i \\
&\leq \frac{9}{8} y_i^2 + \mathbb{1}(y_i < 2^{1-s}) 2^{1-s} y_i \\
&\leq \frac{9}{8} y_i^2 + 2^{1-s} \min(y_i, 2^{1-s}) \\
&= \frac{9 w_i^2}{8\|w\|^2} + 2^{1-s} \min\Big(\frac{|w_i|}{\|w\|}, 2^{1-s}\Big).
\end{aligned}
$$

Using this result, we have

$$
\begin{aligned}
\mathbb{E}[\|Q^s_{nat}(w)\|^2] &\leq \sum_{i=1}^{d} \mathbb{E}\big[\|w\|^2 \xi_i(y_i)^2\big] \\
&\leq \|w\|^2 \sum_{i=1}^{d} \left(\frac{9 w_i^2}{8\|w\|^2} + 2^{1-s} \min\Big(\frac{|w_i|}{\|w\|}, 2^{1-s}\Big) \right) \\
&\leq \|w\|^2 \left(\frac{9}{8} + 2^{1-s} \min\Big(\frac{\|w\|_1}{\|w\|}, 2^{1-s} d\Big) \right) \\
&\leq \|w\|^2 \left(\frac{9}{8} + 2^{1-s} \min(\sqrt{d}, 2^{1-s} d \right),
\end{aligned}
$$

where last inequality comes from the fact $\max_w \frac{\|w\|_1}{\|w\|} = \sqrt{d}$. □

3.2 Assumptions

Before going deeper into theoretical results, we will make some common assumptions about our problem, which will be essential for our analysis.

Assumption 3.1 (L-smooth)**.** *Each $f_i : \mathbb{R}^d \to \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$, is L-smooth, i.e., there exists $L > 0$ such that*

$$\|\nabla f_i(w) - \nabla f_i(w')\| \leq L\|w - w'\|, \ \forall w, w' \in \mathbb{R}^d.$$

This assumption implies that $F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ is also L-smooth.

Assumption 3.2 (μ-strongly convex)**.** *$F : \mathbb{R}^d \to \mathbb{R}$, is μ-strongly convex, i.e., there exists $\mu > 0$ such that*

$$F(w) \geq F(w') + \nabla F(w')^T(w - w') + \frac{\mu}{2}\|w - w'\|^2, \ \forall w, w' \in \mathbb{R}^d.$$

Note that this assumption implies

$$\mu\|w - w'\| \leq \|\nabla F(w) - \nabla F(w')\|. \tag{3.1}$$

Assumption 3.3. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$, is strongly convex with $\mu > 0$.*

Assumption 3.4. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$, $\forall i \in \{1, \ldots, n\}$, is convex, i.e.,*

$$f_i(w) \geq f_i(w') + \nabla f_i(w')^T(w - w'), \ \forall w, w' \in \mathbb{R}^d.$$

We will use Assumption 3.1 on all parts of the analysis. For convexity and strong convexity assumptions, we will use them if needed.

3.3 Additional Lemmas

We will use some common results from the assumptions we made from the previous section.

Lemma 3.3 (Theorem 2.1.5 in [8]). *Suppose that f is convex and L-smooth. Then, for any $w, w' \in \mathbb{R}^d$,*

$$f(w) \leq f(w') + \nabla f(w')^T(w - w') + \frac{L}{2}\|w - w'\|^2, \tag{3.2}$$

$$f(w) \geq f(w') + \nabla f(w')^T(w - w') + \frac{1}{2L}\|\nabla f(w) - \nabla f(w')\|^2, \tag{3.3}$$

$$(\nabla f(w) - \nabla f(w'))^T(w - w') \geq \frac{1}{L}\|\nabla f(w) - \nabla f(w')\|^2. \tag{3.4}$$

Note that (3.2) does not require the convexity of f.

Lemma 3.4 (Theorem 2.1.10 in [8]). *Suppose that f is μ-strongly convex. Then, for any $w, w' \in \mathbb{R}^d$,*

$$f(w) \leq f(w') + \nabla f(w')^T(w - w') + \frac{1}{2\mu}\|\nabla f(w) - \nabla f(w')\|^2. \tag{3.5}$$

Note that Lemma 3.4 implies

$$2\mu[f(w) - f(w_*)] \leq \|\nabla f(w)\|^2, \tag{3.6}$$

since ∇f is zero vector at w_*, unique minimizer of f.

Lemma 3.5 (Theorem 2.1.12 in [8]). *Suppose that f is μ-strongly convex and L-smooth. Then, for any $w, w' \in \mathbb{R}^d$,*

$$(\nabla f(w) - \nabla f(w'))^T(w - w') \geq \frac{\mu L}{\mu + L}\|w - w'\|^2 + \frac{1}{\mu + L}\|\nabla f(w) - \nabla f(w')\|^2. \tag{3.7}$$

3.4 QSARAH

We will start our analysis with three important lemmas, like SARAH paper[9]. The first two of these lemmas do not require convexity assumptions. The first lemma

bounds sum of the expected values of gradients in one inner loop, which does not require any modifications to the original lemma.

Lemma 3.6 (Lemma 1 in [9]). *Suppose that Assumption 3.1 holds. Then, we have*

$$\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \le \frac{2}{\eta}\mathbb{E}[F(w_0^{(s)}) - F(w_*)] + \sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \\ -(1-L\eta)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2]. \tag{3.8}$$

Proof. By (2.5) and Assumption 3.1, we have

$$\begin{aligned}\mathbb{E}[F(w_{t+1}^{(s)})] &\overset{(3.2)}{\le} \mathbb{E}[F(w_t^{(s)})] - \eta\mathbb{E}[\nabla F(w_t^{(s)})^T v_t^{(s)}] + \frac{L\eta^2}{2}\mathbb{E}[\|v_t^{(s)}\|^2] \\ &= \mathbb{E}[F(w_t^{(s)})] - \frac{\eta}{2}\mathbb{E}[\|F(w_t^{(s)})\|^2] + \frac{\eta}{2}\mathbb{E}[\|F(w_t^{(s)}) - v_t^{(s)}\|^2] - \Big(\frac{\eta - L\eta^2}{2}\Big)\mathbb{E}[\|v_t^{(s)}\|^2],\end{aligned}$$

where the last equality follows from the fact $a^T b = \frac{1}{2}[\|a\|^2 + \|b\|^2 - \|a-b\|^2]$. By summing over $t = 0, 1, \ldots, m$, we have

$$\mathbb{E}[F(w_{m+1}^{(s)})] \le \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2}\sum_{t=0}^{m}\mathbb{E}[\|F(w_t^{(s)})\|^2] + \frac{\eta}{2}\sum_{t=0}^{m}\mathbb{E}[\|F(w_t^{(s)}) - v_t^{(s)}\|^2] \\ - \Big(\frac{\eta - L\eta^2}{2}\Big)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2],$$

which is equivalent to ($\eta > 0$),

$$\begin{aligned}\sum_{t=0}^{m}\mathbb{E}[\|F(w_t^{(s)})\|^2] &\le \frac{2}{\eta}\mathbb{E}[F(w_0^{(s)}) - F(w_{m+1}^{(s)})] + \sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \\ &\quad - (1-L\eta)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2] \\ &\le \frac{2}{\eta}\mathbb{E}[F(w_0^{(s)}) - F(w_*)] + \sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \\ &\quad - (1-L\eta)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2],\end{aligned}$$

where w_* is global minimizer of F. $\square$

Our second important lemma bounds expected value of $\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2$, without convexity assumption.

Lemma 3.7. *Suppose that Assumption 3.1 holds. Then for any $t \geq 1$*

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] = \sum_{j=1}^{t} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^{t} \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2]. \tag{3.9}$$

Proof. For $j \geq 1$ we have

$$\begin{aligned}
&\mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}[\|\nabla F(w_j^{(s)}) - v_j^{(s)}\|^2] \\
&= \mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}\left[\|[\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)}] + [\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})] - [v_j^{(s)} - v_{j-1}^{(s)}]\|^2\right] \\
&= \|\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)}\|^2 + \|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2 + \mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] \\
&+ 2(\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)})^T(\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})) \\
&- 2(\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)})^T \mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}[v_j^{(s)} - v_{j-1}^{(s)}] \\
&- 2(\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)}))^T \mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}[v_j^{(s)} - v_{j-1}^{(s)}],
\end{aligned}$$

and from (2.4),

$$\begin{aligned}
\mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}[v_j^{(s)} - v_{j-1}^{(s)}] &= \mathbb{E}_{I_j^{(s)}, Q_j^{(s)}}\Big[\frac{1}{b}\sum_{i \in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big] \\
&\overset{(2.1)}{=} \mathbb{E}_{I_j^{(s)}}\Big[\frac{1}{b}\sum_{i \in I_j^{(s)}} \nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\Big] \\
&= \nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)}).
\end{aligned}$$

Substituting second equation into first equation, we have

$$\mathbb{E}_{I_j^{(s)},Q_j^{(s)}}[\|\nabla F(w_j^{(s)}) - v_j^{(s)}\|^2] = \|\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)}\|^2 - \|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2 + \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2],$$

and taking total expectation,

$$\mathbb{E}[\|\nabla F(w_j^{(s)}) - v_j^{(s)}\|^2] = \mathbb{E}[\|\nabla F(w_{j-1}^{(s)}) - v_{j-1}^{(s)}\|^2] - \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2] + \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2].$$

Note that $\nabla F(w_0^{(s)}) = v_0^{(s)}$. By summing over $j = 1, \ldots, t$ $(t \geq 1)$,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] = \sum_{j=1}^{t} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^{t} \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2].$$

□

The last lemma bounds expected value of $\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2$, like Lemma 3.7, but this time making use of convexity.

Lemma 3.8. *Suppose that Assumptions 3.1 and 3.4 hold and $\eta \leq 2/L$. Then for any $t \geq 1$*

$$\begin{aligned}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] &\leq \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)} \left[\mathbb{E}[\|v_0\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right] \\ &\leq \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)} \mathbb{E}[\|v_0^{(s)}\|^2]. \end{aligned} \tag{3.10}$$

Proof. For $j \geq 1$ we have

$$\mathbb{E}_{I_j^{(s)},Q_j^{(s)}}[\|v_j^{(s)}\|^2] \overset{(2.4)}{=} \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\|v_{j-1}^{(s)} + \frac{1}{b}\sum_{i \in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big]$$

$$
\begin{aligned}
&= \|v_{j-1}^{(s)}\|^2 + \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big\|^2\Big] \\
&\quad + 2(v_{j-1}^{(s)})^T \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big] \\
&\overset{(2.1)}{=} \|v_{j-1}^{(s)}\|^2 + \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big\|^2\Big] \\
&\quad + 2(v_{j-1}^{(s)})^T \mathbb{E}_{I_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} \nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\Big] \\
&\overset{(2.5)}{=} \|v_{j-1}^{(s)}\|^2 + \mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big\|^2\Big] \\
&\quad - \frac{2}{\eta}(w_j^{(s)} - w_{j-1}^{(s)})^T \mathbb{E}_{I_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} \nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\Big]. \quad (3.11)
\end{aligned}
$$

If we apply (3.4) to each function in the last term, we have

$$
\begin{aligned}
&(w_j^{(s)} - w_{j-1}^{(s)})^T \mathbb{E}_{I_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} \nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\Big] \\
&\qquad \geq \frac{1}{L}\mathbb{E}_{I_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} \|\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\|\Big] \\
&\qquad \overset{(2.2)}{\geq} \frac{1}{L(1+\omega)}\mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_j^{(s)}} \|Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\|^2\Big] \\
&\qquad \geq \frac{1}{L(1+\omega)}\mathbb{E}_{I_j^{(s)},Q_j^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big\|^2\Big].
\end{aligned}
$$

Combining above inequalities and taking total expectation, we have

$$
\begin{aligned}
\mathbb{E}[\|v_j^{(s)}\|^2] &\leq \mathbb{E}[\|v_{j-1}^{(s)}\|^2] \quad (3.12)\\
&\quad + \Big(1 - \frac{2}{\eta L(1+\omega)}\Big)\mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j^{(s)}} Q_{j,i}^{(s)}(\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)}))\Big\|^2\Big] \\
&\quad (3.13)
\end{aligned}
$$

$$\overset{(2.4)}{=} \mathbb{E}[\|v_{j-1}^{(s)}\|^2] + \left(1 - \frac{2}{\eta L(1+\omega)}\right)\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2],$$

which implies

$$\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] \leq \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)}\Big[\mathbb{E}[\|v_{j-1}^{(s)}\|^2] - \mathbb{E}[\|v_j^{(s)}\|^2]\Big],$$

when $\eta < \frac{2}{L(1+\omega)}$. By summing up over $j = 1, \ldots, t$

$$\sum_{j=1}^{t} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] \leq \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)}\Big[\mathbb{E}[\|v_0\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\Big].$$

By Lemma 3.7, we have

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \sum_{j=1}^{t} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2],$$

combining above two inequalities,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)}\Big[\mathbb{E}[\|v_0\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\Big].$$

□

With these three lemmas, we are ready to present our first result, which is, QSARAH inherits the diminishing step sizes property from SARAH. First, lets look at convex case.

Theorem 3.9. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold and* $\eta < \frac{2}{L(1+\omega)}$. *Then, for any* $t \geq 1$,

$$\begin{aligned}\mathbb{E}[\|v_t^{(s)}\|^2] &\leq \Big[1 - \Big(\tfrac{2}{\eta L(1+\omega)} - 1\Big)\mu^2\eta^2\Big]\mathbb{E}[\|v_{t-1}^{(s)}\|^2] \\ &\leq \Big[1 - \Big(\tfrac{2}{\eta L(1+\omega)} - 1\Big)\mu^2\eta^2\Big]^t \mathbb{E}[\|\nabla F(w_0^{(s)})\|^2].\end{aligned}$$

Proof. For $j \geq 1$ we have

$$\begin{aligned}
\|\nabla F(w_t^{(s)}) - \nabla F(w_{t-1}^{(s)})\|^2 &= \Big\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\Big\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\|^2 \\
&= \mathbb{E}_i[\|(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\|^2] \\
&\leq \mathbb{E}_{i,Q}[\|Q(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\|^2] \\
&= \mathbb{E}_{I_t^{(s)},Q_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}}\|Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\|^2\Big] \\
&\leq \mathbb{E}_{I_t^{(s)},Q_t^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_t^{(s)}}Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\Big\|^2\Big],
\end{aligned}$$

taking expectation,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - \nabla F(w_{t-1}^{(s)})\|^2] \leq \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_t^{(s)}}Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\Big\|^2\Big].$$

Combining (3.13) with the inequality above, since $1 - \frac{2}{\eta L(1+\omega)} < 0$ when $\eta < \frac{2}{L(1+\omega)}$, we have

$$\begin{aligned}
\mathbb{E}[\|v_t^{(s)}\|^2] &\leq \mathbb{E}[\|v_{t-1}^{(s)}\|^2] + \Big(1 - \frac{2}{\eta L(1+\omega)}\Big)\mathbb{E}[\|\nabla F(w_t^{(s)}) - \nabla F(w_{t-1}^{(s)})\|^2] \\
&\overset{(3.1)}{\leq} \mathbb{E}[\|v_{t-1}^{(s)}\|^2] + \Big(1 - \frac{2}{\eta L(1+\omega)}\Big)\mu^2\mathbb{E}[\|w_t^{(s)} - w_{t-1}^{(s)}\|^2] \\
&\overset{(2.5)}{\leq} \mathbb{E}[\|v_{t-1}^{(s)}\|^2] + \Big(1 - \frac{2}{\eta L(1+\omega)}\Big)\mu^2\eta^2\mathbb{E}[\|v_{t-1}^{(s)}\|^2] \\
&= \Big[1 - \Big(\frac{2}{\eta L(1+\omega)} - 1\Big)\mu^2\eta^2\Big]\mathbb{E}[\|v_{t-1}^{(s)}\|^2],
\end{aligned}$$

applying recursively, and using $v_0^{(s)} = \nabla F(w_0^{(s)})$, we have

$$\mathbb{E}[\|v_t^{(s)}\|^2] \leq \Big[1 - \Big(\frac{2}{\eta L(1+\omega)} - 1\Big)\mu^2\eta^2\Big]^t\mathbb{E}[\|v_0^{(s)}\|^2]$$

$$\leq \left[1 - \Big(\frac{2}{\eta L(1+\omega)} - 1\Big)\mu^2\eta^2\right]^t \mathbb{E}[\|\nabla F(w_0^{(s)})\|^2].$$

□

According to Theorem 3.9, $\|v_t^{(s)}\|^2$ linearly converges to zero with the rate $(1 - 1/\kappa^2)$. We can improve this rate for strongly convex case with the following theorem.

Theorem 3.10. *Suppose that Assumptions 3.1 and 3.3 hold and* $\eta < \frac{2}{(\mu+L)(1+\omega)}$. *Then, for any* $t \geq 1$,

$$\begin{aligned}\mathbb{E}[\|v_t^{(s)}\|^2] &\leq \Big(1 - \tfrac{2\mu L\eta}{\mu+L}\Big)\mathbb{E}[\|v_{t-1}^{(s)}\|^2] \\ &\leq \Big(1 - \tfrac{2\mu L\eta}{\mu+L}\Big)^t \mathbb{E}[\|\nabla F(w_0^{(s)})\|^2].\end{aligned}$$

Proof. Starting from (3.11), for $t \geq 1$, we have

$$\begin{aligned}\mathbb{E}_{I_t^{(s)},Q_t^{(s)}}[\|v_t^{(s)}\|^2] &= \|v_{t-1}^{(s)}\|^2 + \mathbb{E}_{I_t^{(s)},Q_t^{(s)}}\Big[\Big\|\frac{1}{b}\sum_{i\in I_t^{(s)}} Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\Big\|^2\Big] \\ &\quad - \frac{2}{\eta}(w_t^{(s)} - w_{t-1}^{(s)})^T\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\Big] \\ &\leq \|v_{t-1}^{(s)}\|^2 + \mathbb{E}_{I_t^{(s)},Q_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|Q_{t,i}^{(s)}(\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)}))\|^2\Big] \\ &\quad - \frac{2}{\eta}(w_t^{(s)} - w_{t-1}^{(s)})^T\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\Big] \\ &\overset{(2.2)}{\leq} \|v_{t-1}^{(s)}\|^2 + (1+\omega)\mathbb{E}_{I_t^{(s)},Q_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\|^2\Big] \\ &\quad - \frac{2}{\eta}(w_t^{(s)} - w_{t-1}^{(s)})^T\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\Big] \\ &\overset{(3.7)}{\leq} \|v_{t-1}^{(s)}\|^2 + (1+\omega)\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\|^2\Big] \\ &\quad - \frac{2}{\eta}\Big[\frac{\mu L}{\mu+L}\|w_t^{(s)} - w_{t-1}^{(s)}\|^2\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{\mu + L}\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\|^2\Big]\Big] \\
&= \|v_{t-1}^{(s)}\|^2 - \frac{2\mu L}{\eta(\mu+L)}\|w_t^{(s)} - w_{t-1}^{(s)}\|^2 \\
&\quad + \left(1 + \omega - \frac{2}{\eta(\mu+L)}\right)\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\|^2\Big] \\
&\overset{(2.5)}{=} \left(1 - \frac{2\mu L\eta}{\mu + L}\right)\|v_{t-1}^{(s)}\|^2 \\
&\quad + \left(1 + \omega - \frac{2}{\eta(\mu+L)}\right)\mathbb{E}_{I_t^{(s)}}\Big[\frac{1}{b}\sum_{i\in I_t^{(s)}} \|\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})\|^2\Big] \\
&\leq \left(1 - \frac{2\mu L\eta}{\mu + L}\right)\|v_{t-1}^{(s)}\|^2,
\end{aligned}
$$

last inequality comes from $(1+\omega-\frac{2}{\eta(\mu+L)}) < 0$, when we set $\eta < \frac{2}{(\mu+L)(1+\omega)}$. Applying recursively, and using $v_0^{(s)} = \nabla F(w_0^{(s)})$, we have

$$
\begin{aligned}
\mathbb{E}[\|v_t^{(s)}\|^2] &\leq \left(1 - \frac{2\mu L\eta}{\mu + L}\right)^t \mathbb{E}[\|v_0^{(s)}\|^2] \\
&\leq \left(1 - \frac{2\mu L\eta}{\mu + L}\right)^t \mathbb{E}[\|\nabla F(w_0^{(s)})\|^2].
\end{aligned}
$$

□

According to Theorem 3.10, $\|v_t^{(s)}\|^2$ linearly converges to zero with the rate $(1 - 1/\kappa)$. Note that this result is much better compared to convex case.

Now we turn our attention into convergence in multiple inner loops. Before doing that. instead of diminishing step sizes, we need to prove that the gradients of iterates are converging to zero with increasing m.

Theorem 3.11. *Suppose that Assumptions 3.1 and 3.4 hold and* $\eta < \frac{1}{L(1+\omega)}$. *Then, for any* $s \geq 1$,

$$
\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \leq \frac{2}{\eta(m+1)}\mathbb{E}[F(\tilde{w}_{s-1}) - F(w_*)] + \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)}\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2].
$$

Proof. By Lemma 3.8, and since $v_0^{(s)} = \nabla F(w_0^{(s)})$, we have

$$\sum_{t=0}^{m} \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \frac{m\eta L(1+\omega)}{2 - \eta L(1+\omega)} \mathbb{E}[\|v_0^{(s)}\|^2],$$

combining with the Lemma 3.6, since $\eta \leq \frac{1}{L(1+\omega)}$, we have

$$\sum_{t=0}^{m} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \frac{2}{\eta} \mathbb{E}[F(w_0^{(s)}) - F(w_*)] + \frac{m\eta L(1+\omega)}{2 - \eta L(1+\omega)} \mathbb{E}[\|v_0^{(s)}\|^2]. \tag{3.14}$$

Note that above inequality is for single outer iteration. Also note that $w_0^{(s+1)} = w_{\bar{t}}^{(s)}$, with $s \geq 0$, where $\bar{t}$ is picked uniformly random from $\{0, 1, \ldots, m\}$. Hence, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla F(w_0^{(s+1)})\|^2] &= \frac{1}{m+1} \sum_{t=0}^{m} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
&\overset{(3.14)}{\leq} \frac{2}{\eta(m+1)} \mathbb{E}[F(w_0^{(s)}) - F(w_*)] + \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)} \frac{m}{m+1} \mathbb{E}[\|v_0^{(s)}\|^2] \\
&\leq \frac{2}{\eta(m+1)} \mathbb{E}[F(w_0^{(s)}) - F(w_*)] + \frac{\eta L(1+\omega)}{2 - \eta L(1+\omega)} \mathbb{E}[\|v_0^{(s)}\|^2].
\end{aligned}$$

□

Theorem 3.11 implies sublinear convergence of the gradient with increasing m. With following corollary, we show the exact complexity bound for QSARAH within a single inner loop.

Corollary 3.11.1. *Consider QSARAH (Algorithm 2) within single outer iteration with learning rate* $\eta = \sqrt{\frac{2}{L(m+1)(1+\omega)}}$ *where* $m > 2L(1+\omega) - 1$, *then* $\|\nabla F(w_t^{(s)})\|^2$ *converges sublinearly in expectation with a rate of* $\sqrt{\frac{2L}{(m+1)(1+\omega)}}$ *and have the total complexity of* $\mathcal{O}(n + \frac{1}{\epsilon^2})$.

Proof. Theorem 3.11 with $\eta \leq \frac{1}{L(1+\omega)}$ implies

$$\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \leq \frac{2}{\eta(m+1)} \mathbb{E}[F(\tilde{w}_{s-1}) - F(w_*) + \eta L(1+\omega) \mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2],$$

by choosing learning rate $\eta = \sqrt{\frac{2}{L(m+1)(1+\omega)}}$ and $m > 2L(1+\omega) - 1$ in order to ensure $\eta \leq \frac{1}{L(1+\omega)}$, we have

$$\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \leq \sqrt{\frac{2L}{(m+1)(1+\omega)}}\mathbb{E}[F(\tilde{w}_{s-1}) - F(w_*) + \|\nabla F(\tilde{w}_{s-1})\|^2],$$

which implies the result. □

By using Theorem 3.11, we prove convergence of the gradients for multiple inner loops in the following theorem for convex case.

Theorem 3.12. *Suppose that Assumptions 3.1 and 3.4 hold. Consider QSARAH and define* $\delta_k = \frac{2}{\eta(m+1)}\mathbb{E}[F(\tilde{w}_k) - F(w_*)], k = 0, 1, \ldots, s-1,$ *and* $\delta = \max_{k\in\{0,1,\ldots,s-1\}} \delta_k$. *Then we have*

$$\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] - \Delta \leq \alpha^s(\|\nabla F(\tilde{w}_0)\|^2 - \Delta) \tag{3.15}$$

where $\alpha = \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)}$, *and* $\Delta = \frac{\delta}{1-\alpha}$.

Proof. By Theorem 3.11, we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] &\leq \frac{2}{\eta(m+1)}\mathbb{E}[F(\tilde{w}_{s-1}) - F(w_*)] + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)}\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2] \\ &\leq \delta_{s-1} + \alpha\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2],\end{aligned}$$

applying recursively,

$$\begin{aligned}\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] &\leq \delta_{s-1} + \alpha\delta_{s-2} + \cdots + \alpha^{s-1}\delta_0 + \alpha^s\|\nabla F(\tilde{w}_0)\|^2 \\ &\leq \delta + \alpha\delta + \cdots + \alpha^{s-1}\delta + \alpha^s\mathbb{E}[\|\nabla F(\tilde{w}_0)\|^2 \\ &\leq \delta\frac{1-\alpha^s}{1-\alpha} + \alpha^s\|\nabla F(\tilde{w}_0)\|^2 \\ &= \Delta(1-\alpha^s) + \alpha^s\|\nabla F(\tilde{w}_0)\|^2\end{aligned}$$

$$= \Delta + \alpha^s(\|\nabla F(\tilde{w}_0)\|^2 - \Delta).$$

□

The total complexity to have ϵ-accurate solution can be seen in the following corollary.

Corollary 3.12.1. *Consider QSARAH (Algorithm 2). Suppose we choose* $\Delta = \frac{\epsilon}{4}$, $\alpha = \frac{1}{2}$ *(with* $\eta = \frac{2}{3L(1+\omega)}$*), and* $m = \mathcal{O}(\frac{1}{\epsilon})$ *in Theorem 3.12. Then, the total complexity to achieve* ϵ*-accurate solution is* $\mathcal{O}((n + \frac{1}{\epsilon})\log(\frac{1}{\epsilon}))$.

Proof. Note that, in Theorem 3.12, in order to have ϵ-accurate solution, Δ needs to be less than ϵ, which can happen if $m = \frac{\infty}{\epsilon}$. Writing Theorem 3.12 with constants given in Corollary 3.12.1, we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] &\leq \Delta + \alpha^s(\|\nabla F(\tilde{w}_0)\|^2 - \Delta) \\ &\leq \Delta + \alpha^s\|\nabla F(\tilde{w}_0)\|^2 \\ &= \frac{\epsilon}{4} + \frac{1}{2^s}\|\nabla F(\tilde{w}_0)\|^2,\end{aligned}$$

to have $\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \leq \epsilon$, $\frac{1}{2^s}\|\nabla F(\tilde{w}_0)\|^2 \leq \frac{3}{4}\epsilon$, which means $s = \mathcal{O}(\log(\frac{1}{\epsilon}))$. We have complexity of $n + 2m$ for each inner loop, hence the total complexity to achieve ϵ-accurate solution is $\mathcal{O}((n + \frac{1}{\epsilon})\log(\frac{1}{\epsilon}))$. □

For the strongly convex case, again by using Theorem 3.11, we prove convergence of the gradients for multiple inner loops in the following theorem for strongly convex case.

Theorem 3.13. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Consider QSARAH with the choice of* η *and* m*, define* $\sigma = \frac{1}{\mu\eta(m+1)} + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)} < 1$*. Then we have*

$$\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] \leq \sigma^s\|F(\tilde{w}_0)\|^2. \tag{3.16}$$

Proof. By Theorem 3.11, we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] &\leq \frac{2}{\eta(m+1)}\mathbb{E}[F(\tilde{w}_{s-1}) - F(w_*)] + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)}\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2] \\ &\overset{3.6}{\leq} \frac{1}{\mu\eta(m+1)}\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2] + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)}\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2] \\ &= \left(\frac{1}{\mu\eta(m+1)} + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)}\right)\mathbb{E}[\|\nabla F(\tilde{w}_{s-1})\|^2],\end{aligned}$$

applying recursively, we have desired result. □

The total complexity to have ϵ-accurate solution can be seen in the following corollary.

Corollary 3.13.1. *Consider QSARAH (Algorithm 2). Suppose we choose* $\eta = \frac{1}{2L(1+\omega)}$, $m = 4.5\frac{L(1+\omega)}{\mu} = 4.5\kappa(1+\omega)$ *and* $s = \lceil \log(\|\nabla F(\tilde{w}_0)\|^2/\epsilon)/\log(9/7)\rceil$ *in Theorem 3.13. Then, the total complexity to achieve* ϵ*-accurate solution is* $\mathcal{O}((n + \kappa(1+\omega))\log(\frac{1}{\epsilon}))$.

Proof. We can calculate σ in Theorem 3.13 with constants given in Corollary 3.13.1 as

$$\begin{aligned}\sigma &= \frac{1}{\mu\eta(m+1)} + \frac{\eta L(1+\omega)}{2-\eta L(1+\omega)} \\ &= \frac{1}{\frac{\mu}{2L(1+\omega)}(4.5\kappa(1+\omega)+1)} + \frac{\frac{1}{2}}{2-\frac{1}{2}} \\ &< \frac{4}{9} + \frac{1}{3} = \frac{7}{9}.\end{aligned}$$

Using this inequality and Theorem 3.13, and since

$$\begin{aligned}s &= \lceil \log(\|\nabla F(\tilde{w}_0)\|^2/\epsilon)/\log(9/7)\rceil \\ &\geq \log_{7/9}(\epsilon/\|F(\tilde{w}_0)\|^2),\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\tilde{w}_s)\|^2] &\leq \left(\frac{7}{9}\right)^s \|F(\tilde{w}_0)\|^2 \\ &\leq \left(\frac{7}{9}\right)^{\log_{7/9}(\epsilon/\|F(\tilde{w}_0)\|^2)} \|F(\tilde{w}_0)\|^2 = \epsilon.\end{aligned}$$

Since we have complexity of $n + 2m$ for each inner loop, the total complexity to achieve ϵ-accurate solution is $\mathcal{O}((n + \kappa(1 + \omega)) \log(\frac{1}{\epsilon}))$. □

Chapter 4

Experimental Results

In our experiments, we used two datasets in order to compare QSARAH+ with vanilla SARAH, which are MNIST and CIFAR10 datasets. The number of samples and coordinates can be seen on Table 4.1. Our test problem is linear classifier with cross entropy loss at the end, which is a convex problem. We used QSARAH+ with no quantization (or identity operator as quantization) to compare our results instead of vanilla SARAH+, which reduces to vanilla SARAH+ with mini-batches.

Table 4.1: Number of coordinates and samples of the datasets used for experiments.

Dataset	**d**	**n**	**Classes**
MNIST	784	60,000	10
CIFAR10	3,072	50,000	10

In our experiments, we used 10 nodes, and each node chooses 10 samples from the data it has in each inner loop iteration. Note that we divided the whole dataset into 10 partitions for each node beforehand and each node has only access to its partition throughout learning. Also note that this partitioning is kept for different experiment types, in other words, each node has access to the same subset of data for all experiments with the same dataset. We set $\gamma = 0.125$ and $m = 0.05n$ for all our experiments.

Our first experiment is conducted to find the optimal learning rate for QSARAH+ with no quantization. We will use the results with best learning rate in this experiment as a baseline in order to compare QSARAH+ with different quantization schemes. The results of this experiment can be seen on Figure 4.1.

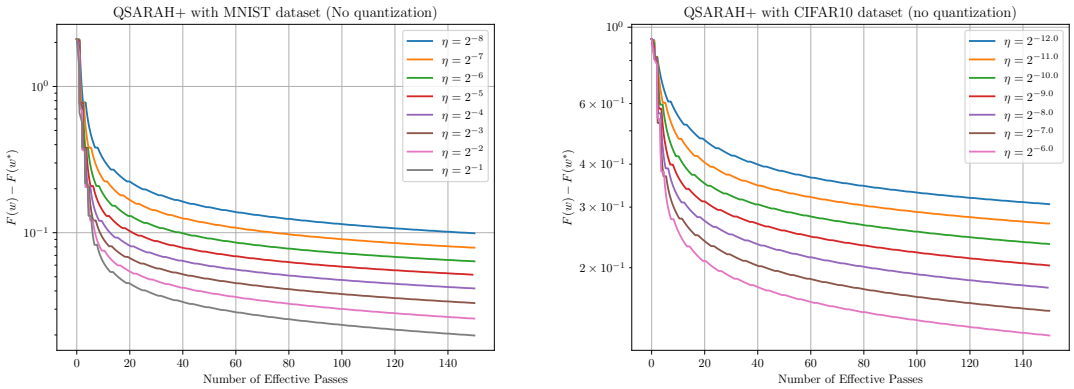

Figure 4.1: Comparison of loss residuals $F(w) - F(w_*)$ of QSARAH+ iterates for MNIST (left) and CIFAR10 (right) datasets with different learning rates and no quantization

We also conducted a series of experiments in order to find the best learning rates for different quantization schemes. Remember that, bits required for s levels of quantization per coordinate can be calculated as $\lceil \log_2(s+1) \rceil + 1$. As a result of this formula, some of the quantization schemes require the same number of bits to communicate. For example, quantization with $s = 9, 10, \ldots, 15$ yield same number of bits per coordinate, which is 5. This means using $s = 9, 10, \ldots, 14$ for this example does not make sense, since reducing s introduces additional variance without any communication savings. For this reason, we used $s = 1, 3, 7, 15, 31, 63$ in our experiments, which corresponds to $2, 3, 4, 5, 6, 7$ bits to communicate per coordinate respectively. The results of QSARAH+ on MNIST dataset can be seen on Figure 4.2 and Figure 4.3, with uniform quantization and natural quantization respectively. The results of QSARAH+ on CIFAR10 dataset can also be seen on Figure 4.4 and Figure 4.5, with uniform quantization and natural quantization respectively.

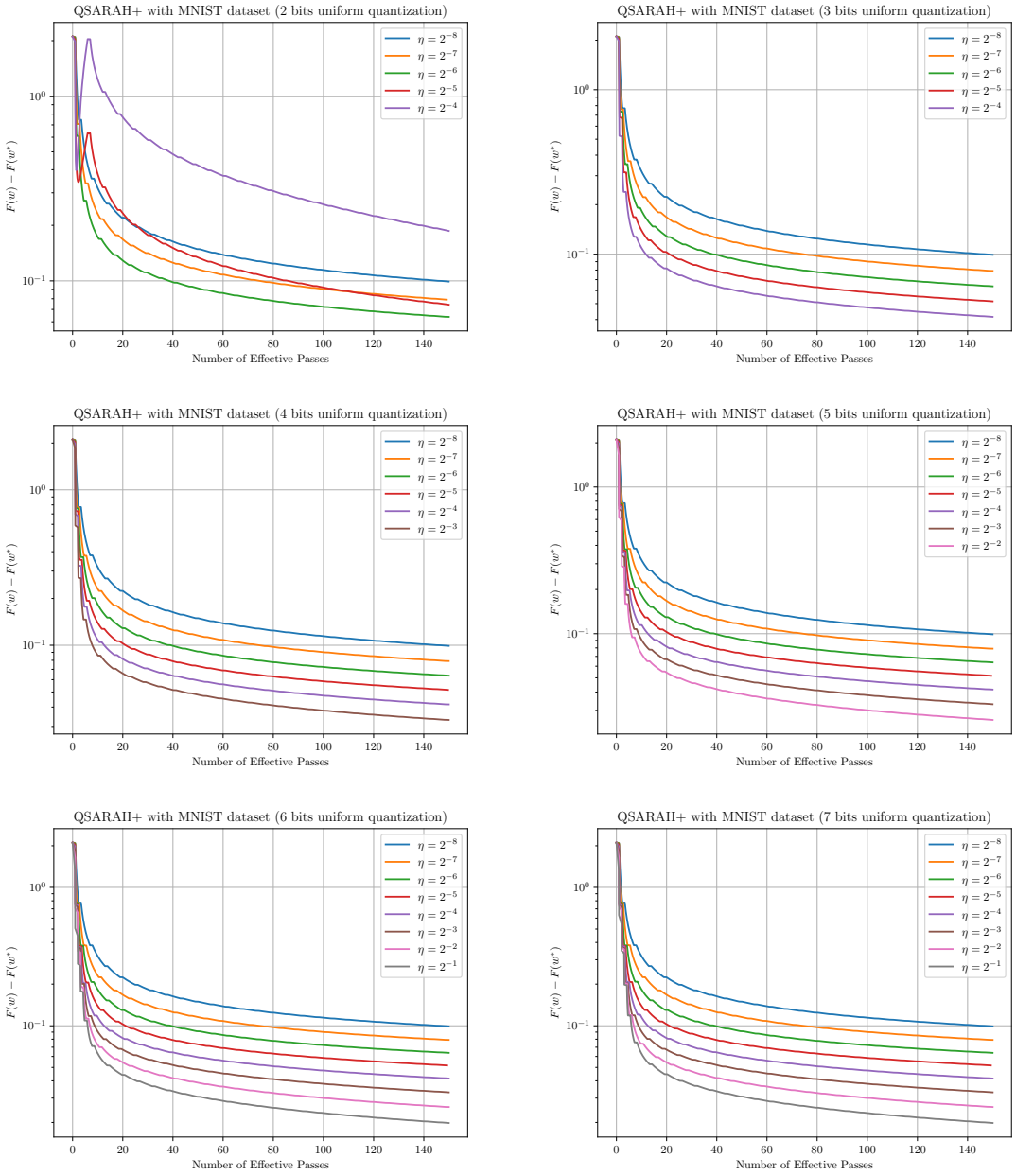

Figure 4.2: Comparison of loss residuals $F(w) - F(w_*)$ of QSARAH+ iterates for MNIST dataset with different learning rates and uniform quantization

We now know optimum learning rates for each quantization scheme, which can be seen in Table 4.2. In order to find the best bit rates on each dataset for uniform and natural quantization, we made plots of the number of bits communicated versus loss residuals. These plots can be seen on Figure 4.6 and Figure 4.7, for MNIST and CIFAR10 datasets respectively.

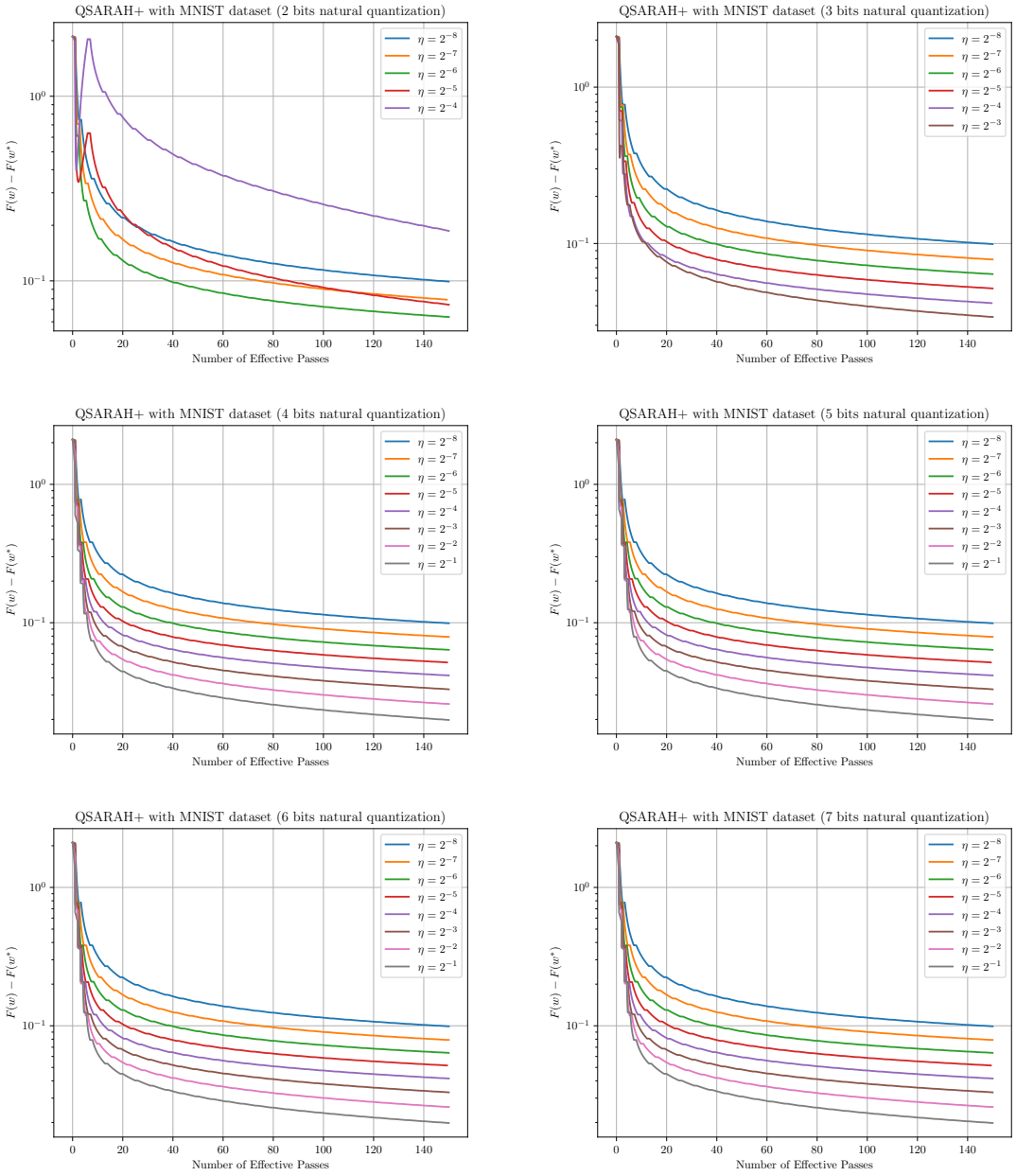

Figure 4.3: Comparison of loss residuals $F(w) - F(w_*)$ of QSARAH+ iterates for MNIST dataset with different learning rates and natural quantization

Our final experiment is comparing best bitrate-learning rate combinations of uniform and natural quantization with no quantization. It can be seen from Figure 4.8 that both uniform and natural quantization schemes can save communication costs greatly.

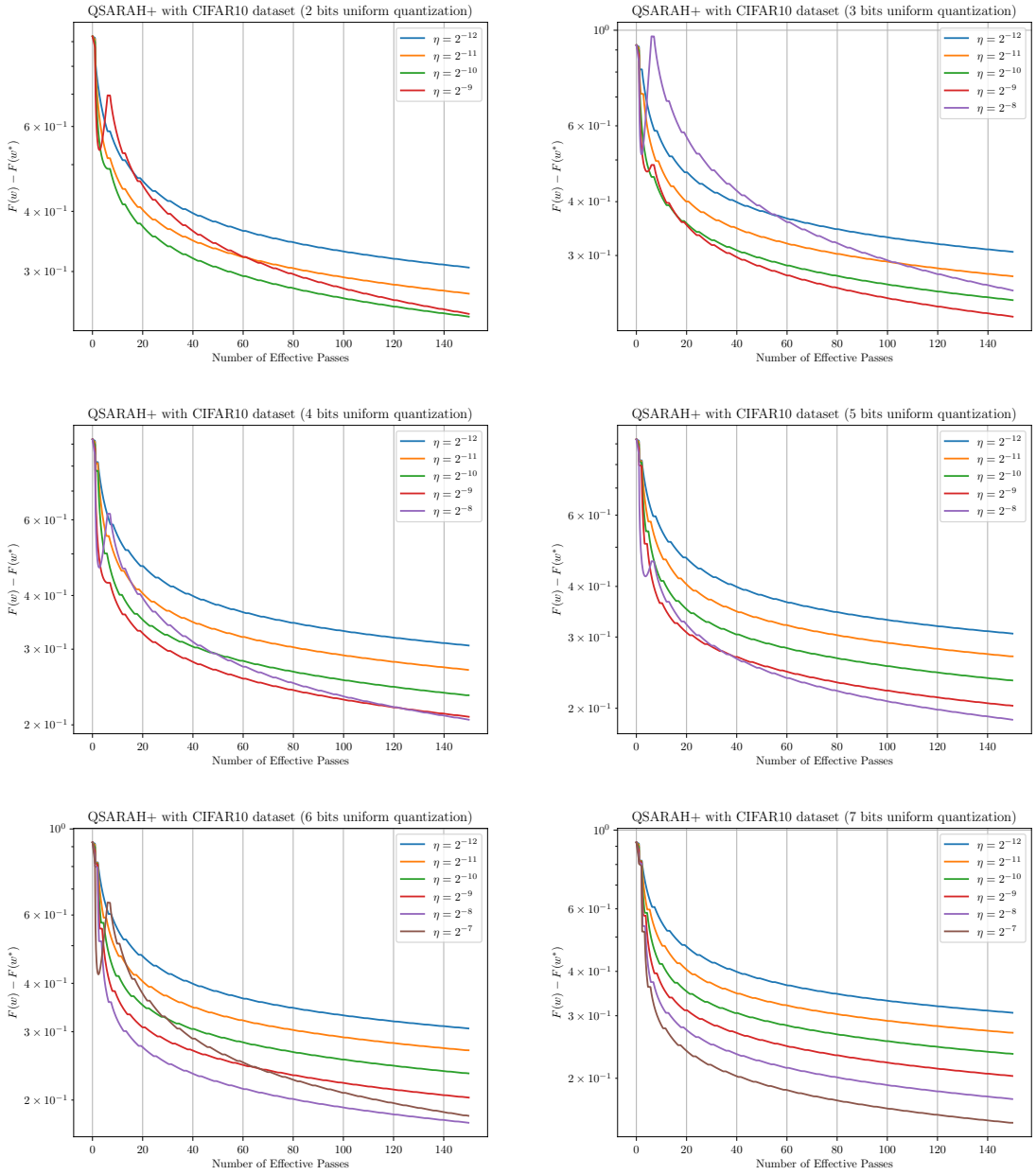

Figure 4.4: Comparison of loss residuals $F(w) - F(w_*)$ of QSARAH+ iterates for CIFAR10 dataset with different learning rates and uniform quantization

Table 4.2: Optimal learning rates of QSARAH+ with different quantization schemes

Dataset	**No qt.**	**Uniform quantization**						**Natural quantization**					
		2b	3b	4b	5b	6b	7b	2b	3b	4b	5b	6b	7b
MNIST	2^{-1}	2^{-6}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{-1}	2^{-6}	2^{-3}	2^{-1}	2^{-1}	2^{-1}	2^{-1}
CIFAR10	2^{-6}	2^{-10}	2^{-10}	2^{-9}	2^{-9}	2^{-8}	2^{-7}	2^{-10}	2^{-9}	2^{-7}	2^{-6}	2^{-6}	2^{-6}

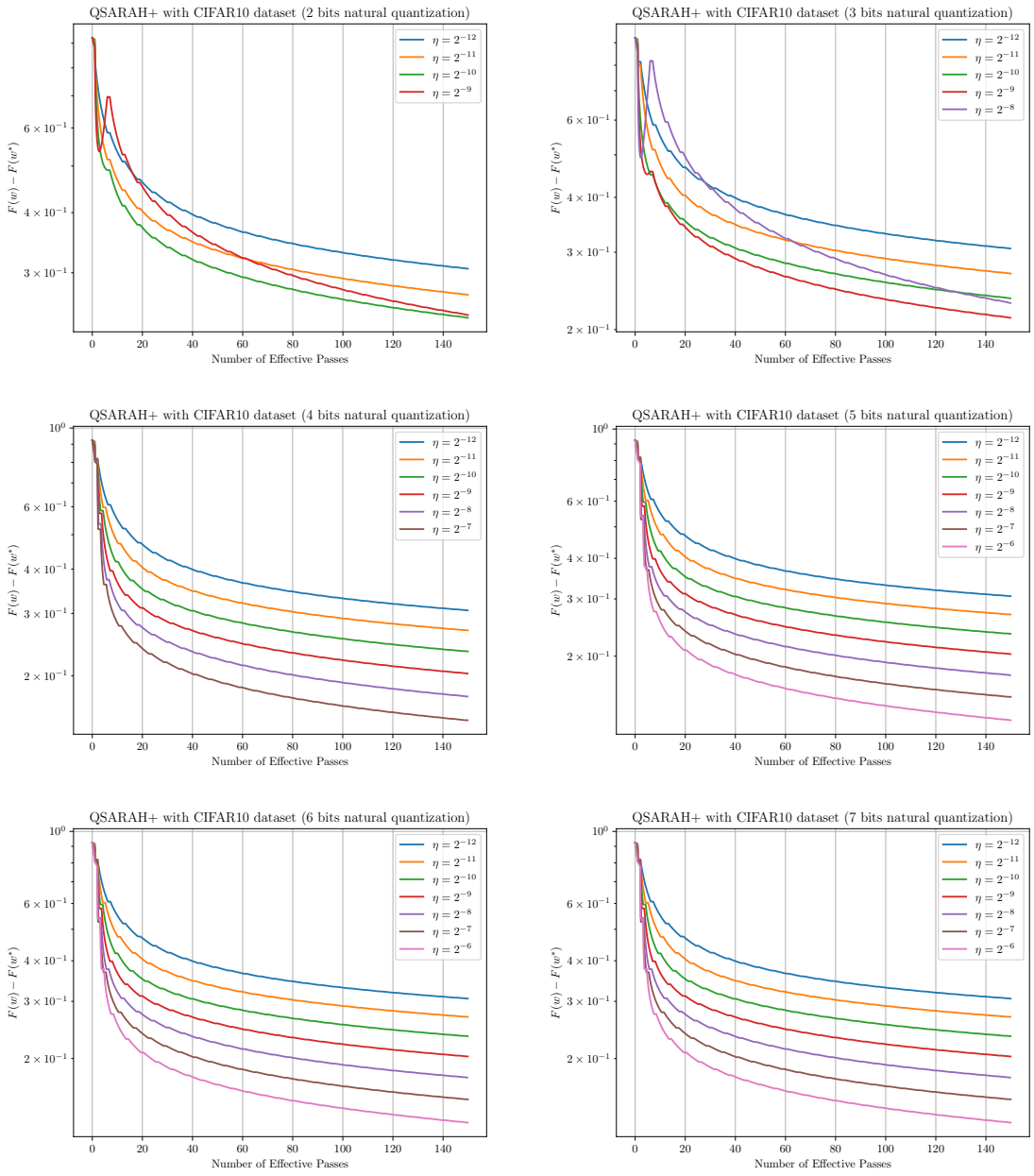

Figure 4.5: Comparison of loss residuals $F(w) - F(w_*)$ of QSARAH+ iterates for CIFAR10 dataset with different learning rates and natural quantization levels

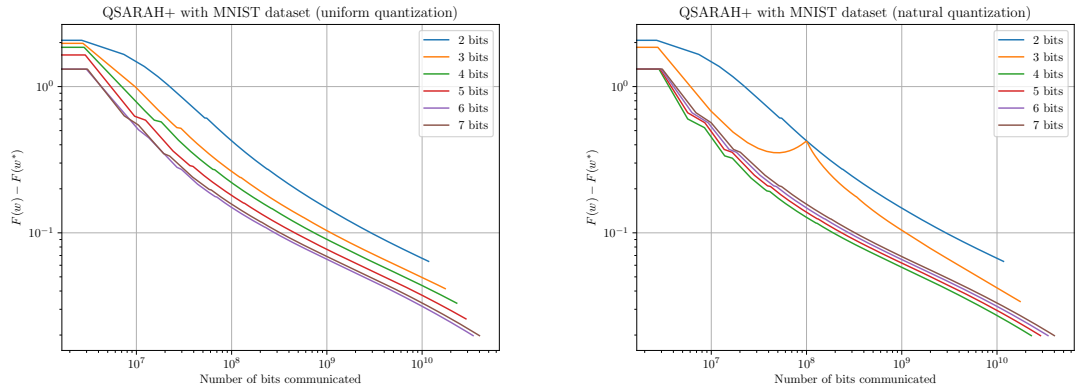

Figure 4.6: Communication cost vs loss residuals $F(w) - F(w_*)$ of QSARAH+ with uniform and natural quantizations for MNIST dataset

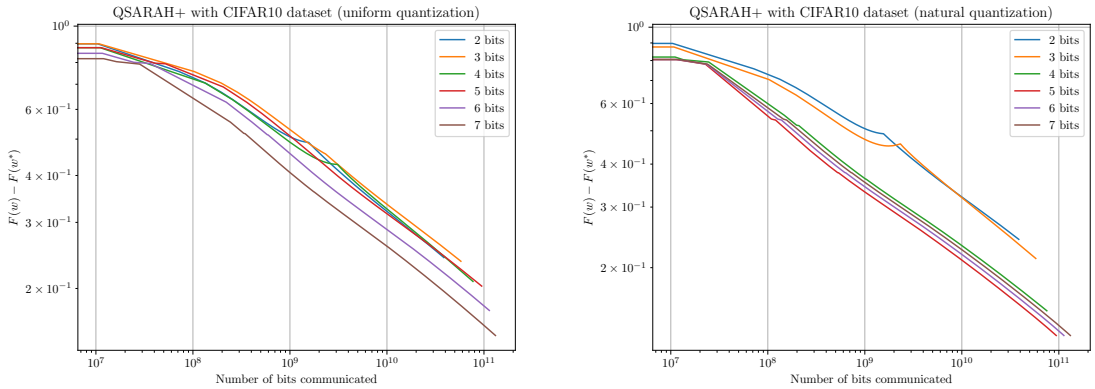

Figure 4.7: Communication cost vs loss residuals $F(w) - F(w_*)$ of QSARAH+ with uniform and natural quantizations for CIFAR10 dataset

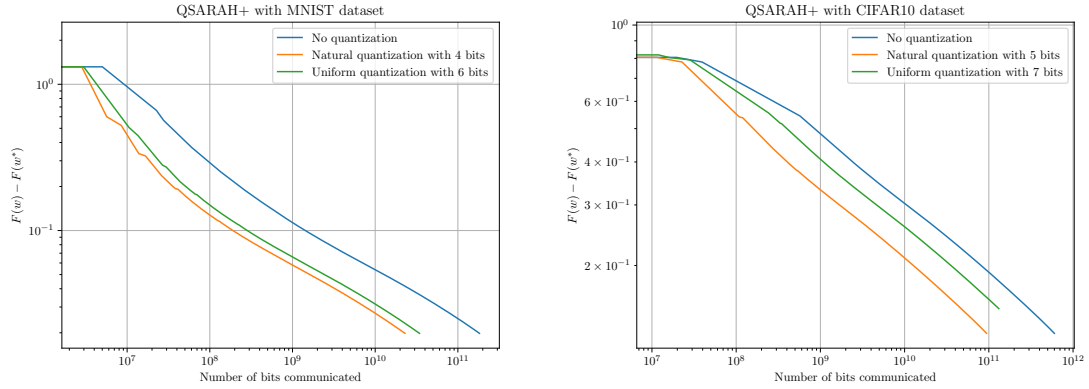

Figure 4.8: Communication cost vs loss residuals $F(w) - F(w_*)$ of QSARAH+ with uniform and natural quantizations on MNIST(left) and CIFAR10(right) datasets

Chapter 5

Conclusion

In this thesis, we introduced a new variant of the SARAH algorithm which is called Quantized SARAH (QSARAH) for empirical risk minimization problem. Our proposed algorithm is suitable for the federated learning setting. Quantization methods used in QSARAH can also be seen as an additional level of security for privacy concerns of federated learning. We introduced two quantization methods, uniform and natural quantization. We also made sure these two methods have some stochastic properties that are crucial for our theoretical analysis. In our theoretical analysis, we showed that QSARAH has diminishing step sizes property in the inner loop like SARAH, which helps with choosing inner loop size. Also, we proved that QSARAH converges to optimal solution linearly for convex and strongly convex problems. Our experimental results are supporting our theoretical analysis and we showed that our algorithm reduces communication costs greatly and achieved lower objective value with less communication.

Bibliography

[1] Dan Alistarh, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: randomized quantization for communication-optimal stochastic gradient descent. *CoRR*, abs/1610.02132, 2016.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives, 2014.

[3] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2017.

[4] Samuel Horvath, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *CoRR*, abs/1905.10988, 2019.

[5] Martin Jaggi, Virginia Smith, Martin Takáč, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I. Jordan. Communication-efficient distributed dual coordinate ascent, 2014.

[6] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[7] Jakub Konečný, Brendan McMahan, and Daniel Ramage. Federated optimization:distributed optimization beyond the datacenter, 2015.

[8] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[9] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient. *arXiv e-prints*, page arXiv:1703.00102, Feb 2017.

[10] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. 17, 10 2013.

[11] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[12] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets, 2012.

[13] Christopher De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild!-style algorithms, 2015.

[14] F. Seide, H. Fu, Jasha Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. pages 1058–1062, 01 2014.

[15] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, mar 2011.

[16] Sebastian U. Stich. Local sgd converges fast and communicates little, 2018.