Distributed learning with compressed gradients

Sarit Khirirat
Automatic Control Department
KTH Royal Institute of Technology
sarit@kth.se

Hamid Reza Feyzmahdavian
ABB Corporate Research Center
Västerås, Sweden
hamid.feyzmahdavian@se.abb.com

Mikael Johansson
Automatic Control Department
KTH Royal Institute of Technology
mikaelj@kth.se

Abstract

Asynchronous computation and gradient compression have emerged as two key techniques for achieving scalability in distributed optimization for large-scale machine learning. This paper presents a unified analysis framework for distributed gradient methods operating with staled and compressed gradients. Non-asymptotic bounds on convergence rates and information exchange are derived for several optimization algorithms. These bounds give explicit expressions for step-sizes and characterize how the amount of asynchrony and the compression accuracy affect iteration and communication complexity guarantees. Numerical results highlight convergence properties of different gradient compression algorithms and confirm that fast convergence under limited information exchange is indeed possible.

1 Introduction

Several problems in machine learning can be cast as empirical risk minimization problems

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) = \sum_{i=1}^{m} f_i(x) \tag{1}$$

where each component function $f_i : \mathbb{R}^d \to \mathbb{R}$ is smooth and has Lipschitz-continuous gradient. The standard first-order method for solving (1) is *gradient descent* (GD)

$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{m} \nabla f_i(x_k) \tag{2}$$

for some positive step-size γ_k. However, when the number of component functions m is extremely large, the computation cost per iteration of GD becomes significant, and one typically resorts to *stochastic gradient descent* (where the gradient is evaluated at a single randomly chosen data point in every iteration) or leverages on data-parallelism by distributing the gradient computations on multiple parallel machines; see *e.g.* in, [1–4]. The latter option leads to *master-server* architectures, where a central master node maintains the current parameter iterate and workers evaluate gradients of the loss on individual subsets of the global data. There are both *synchronous* and *asynchronous* versions of this master-worker architecture.

In the *synchronous* master-worker architecture, the master node waits for all the gradients computed by the workers before it makes an update [3, 5]. Insisting on a synchronous operation leads to long communication times (waiting for the slowest worker to complete) and the benefits of parallelization

Preprint. Work in progress.

diminish as the number of workers increases. *Asynchronous* master-worker architectures, such as parameter server [4], attempt to alleviate this bottleneck by letting the master update its parameters every time it receives new information from a worker. Since the workers now operate on inconsistent data, the training accuracy may degrade and there is a risk that the optimization process diverges.

The natural implementation of distributed gradient descent in the parameter server framework is referred to as *incremental aggregate gradient* (IAG) [6]. Given an initial point x_0 and a step-size γ, the master executes the updates

$$x_{k+1} = x_k - \gamma \sum_{i=1}^{m} \nabla f_i(x_{k-\tau_k^i}). \tag{3}$$

Here τ_k^i describes the staleness of the gradient information from worker i available to the master at iteration k. Under the assumption of bounded staleness, $\tau_k^i \leq \tau$ for all k, i, convergence guarantees for IAG have been established for several classes of loss functions, see *e.g.* [6–9].

A drawback with the master-worker architecture is the massive amount of data exchanged between workers and master. This is especially true when the parameter dimension d is large and we try to scale up the number of worker machines m. Recently, several authors have proposed various gradient compression algorithms for reducing the network cost in distributed machine learning [10–14]. The compression algorithms can be both randomized [10, 11, 13] and deterministic [11, 14], and empirical studies have demonstrated that they can yield significant savings in network traffic [10, 11, 13]. However, the vast majority of the work on gradient compression do not provide convergence guarantees, and the few convergence results that exist often make restrictive assumptions, *e.g.* that component function gradients are uniformly bounded in norm. Even though this assumption is valid for a certain classes of optimization problems, it is always violated when the objective function is strongly convex [15]. In addition, the theoretical support for quantifying the trade-off between iteration and communication complexity is limited, and there are very few general results which allow to characterize the impact of different compression strategies on the convergence rate guarantees.

Contributions. We establish a unified framework for both synchronous and asynchronous distributed optimization using compressed gradients. The framework builds on unbiased randomized quantizers (URQs), a class of gradient compression schemes which cover the ones proposed in [10, 11]. We establish per-iteration convergence rate guarantees for both GD and IAG with URQ compression. The convergence rate guarantees give explicit formulas for how quantization accuracy and staleness bounds affect the expected time to reach an ε-optimal solution. These results allows us to characterize the trade-off between iteration and communication complexity under gradient compression. Finally, we validate the theoretical results on large-scale parameter estimation problems.

我们使用压缩梯度建立了同步和异步分布式优化的统一框架。该框架基于无偏随机量化器(URQs)，这是一类梯度压缩方案，涵盖了[10,11]中提出的方案。通过URQ压缩，我们建立了GD和IAG的每次迭代收敛速度保证。收敛速度保证给出了量化精度和过时界限如何影响预期时间达到“-最优解”的显式公式。这些结果使我们能够描述梯度压缩下迭代和通信复杂性之间的权衡。最后，对大规模参数估计问题的理论结果进行了验证

Related work. Although the initial results on communication complexity of convex optimization appeared over 30 years ago [16], the area has attracted strong renewed interest due to the veritable explosion of data and parameter sizes in deep learning. Several heuristic gradient compression techniques have been proposed and evaluated empirically [10, 13, 17]. Most compression schemes are based on sparsification [10], quantization [13, 14], or combinations of the two [11]; they are either randomized [10, 13] or deterministic [11]. While the majority of papers on gradient compression have a practical focus, several recent works establish theoretical convergence guarantees for gradient compression. In some cases, convergence guarantees are asymptotic, while other papers provide non-asymptotic bounds. The work which is most closely related to the present paper is [11] and [12]. In particular [11] proposes a low-precision quantizer and derives non-asymptotic convergence guarantees for (synchronous) stochastic gradient descent, while [12] introduces an analysis framework based on rate-supermartingales and develops probabilistic guarantees for quantized SGD.

2 Notations and Assumptions

We let $\mathbb{N}, \mathbb{N}_0$ be a set of natural numbers and of natural numbers including zero. For any integers a, b with $a \leq b$, $[a, b] = \{a, a+1, \ldots, b-1, b\}$. For a vector $x \in \mathbb{R}^d$, x^i denotes its i^{th} element, $\text{sign}(x^i)$ the sign of its i^{th} element, and $\text{sign}(x)$ is its sign vector; $\|x\|_0$ denotes the ℓ_0 norm of x or the number of its non-zero elements, $\|x\|$ is its Euclidean norm, and $\text{supp}(x)$ is its support set, i.e.

$$\text{supp}(x) = \{i \mid x^i \neq 0\}.$$

In addition, we impose the following typical assumptions on Problem (1).

Assumption 1. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and has Lipschitz continuous gradient with L, i.e*

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x).y - x\rangle + \frac{L}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

Note that Assumption 1 implies that f also has Lipschitz continuous gradient with $\bar{L} = mL$.

Assumption 2. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mu-$strongly convex, i.e.*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

The Lipschitz constant can be reduced when the gradients of the individual loss functions f_i are sparse. A smaller Lipschitz constant typically translates into larger allowable step-sizes and faster algorithm convergence. Although it is difficicult to quantify the gradient sparsity for general loss functions, it is possible to do so under the following additional assumptions.

Assumption 3. *Each $f_i : \mathbb{R}^d \to \mathbb{R}$ can be written as $f_i(x) = \ell(a_i^T x, b_i)$, such that* $\mathrm{supp}(\nabla f_i(x)) = \mathrm{supp}(a_i)$ *for given data points $\{(a_i, b_i)\}_{i=1}^m$ with $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$.*

Assumption 3, which is satisfied for standard empirical risk minimization problems, implies that the sparsity pattern of gradients can be computed off-line directly from the data. We will consider two important sparsity measures: the average and maximum conflict graph degree of the data, defined as

$$\Delta_{\mathrm{ave}} = \frac{1}{m}\sum_{i=1}^{m}\left\{\sum_{j=1, j\neq i}^{m} \mathbf{1}\{\mathrm{supp}(a_i) \cap \mathrm{supp}(a_j) \neq \emptyset\}\right\}$$

$$\Delta_{\max} = \max_{i\in[1,m]}\left\{\sum_{j=1, j\neq i}^{m} \mathbf{1}\{\mathrm{supp}(a_i) \cap \mathrm{supp}(a_j) \neq \emptyset\}\right\}.$$

As shown next, these sparsity measures allow us to derive a tighter bound $\bar{L}$ for the Lipschitz constant of the total loss:

Lemma 1. *Consider the optimization problem* (1) *under Assumption 3. If ℓ has L-Lipschitz continuous gradient, then the gradient of the total loss is $\bar{L}$-Lipschitz continuous with*

$$\bar{L} = L\sqrt{m(1 + \Delta)},$$

where $\Delta = \min(\Delta_{\mathrm{ave}}, \Delta_{\max})$.

Proof. See Appendix A. □

These sparsity measures are used to tighten our convergence results, especially in Section 5 and 6.

3 Unbiased random quantization

In this paper, we are interested in optimization using unbiased randomized quantizers (URQs):

Definition 1. *A mapping $Q : \mathbb{R}^d \to \mathbb{R}^d$ is called an* unbiased random quantizer *if, for every $v \in \mathbb{R}^d$,*

1. $\mathrm{supp}(Q(v)) \subseteq \mathrm{supp}(v)$
2. $\mathbf{E}\{Q(v)\} = v$
3. $\mathbf{E}\{\|Q(v)\|^2\} \leq \alpha\|v\|^2$

for some finite positive α. In addition, Q is said to be sign-preserving *if*

$$\mathrm{sign}(Q(v)) = \mathrm{sign}(v)$$

for every $v \in \mathbb{R}^d$.

Unbiased random quantizers satisfy some additional useful inequalities. First,

$$\mathbf{E}\left\{\|Q(v)\|_0\right\} \leq c,$$

for any $v \in \mathbb{R}^d$ and a finite positive constant $c \leq d$. If Q is also sign-preserving then

$$\mathbf{E}\,\|Q(v) - v\|^2 \leq \beta\|v\|^2,$$

for any $v \in \mathbb{R}^d$ and a finite positive constant $\beta \leq \alpha - 1$. As we will show next, it is typically possible to derive better bounds for c and β when we consider specific classes of gradient compressors.

3.1 Examples of unbiased random quantizers

Several randomized gradient compression algorithms have been proposed for distributed optimization problems under limited communications. Important examples include the *gradient sparsifier* [10], the *low-precision quantizer* [11] and the *ternary quantizer* [13] defined below.

Definition 2. *The gradient sparsifier $S : \mathbb{R}^d \to \mathbb{R}^d$ is defined as*

$$S^i(v) = \begin{cases} v^i/p^i & \text{with probability } p^i \\ 0 & \text{otherwise} \end{cases},$$

where p^i is probability that coordinate i is selected.

Note that when the gradient sparsifier uses the same probability for each coordinate, it will effectively result in a randomized coordinate descent. Choosing $p^i = |v^i|/\|v\|$, on the other hand, will result in the ternary quantizer [13]:

Definition 3. *The ternary quantizer $T : \mathbb{R}^d \to \mathbb{R}^d$ is defined as*

$$T^i(v) = \begin{cases} \|v\|\text{sign}(v^i) & \text{with probability } |v^i|/\|v\| \\ 0 & \text{otherwise} \end{cases}.$$

The low-precision quantizer [11], defined next, combines sparsification of the gradient vector with quantization of its element to further reduce the amount of information exchanged.

Definition 4. *The low-precision quantizer $Q_b : \mathbb{R}^d \to \mathbb{R}^d$ is defined as*

$$Q_b^i(v) = \|v\|\text{sign}(v^i)\xi(v, i, s),$$

where

$$\xi(v, i, s) = \begin{cases} l/s & \text{with probability } 1 - p\left(|v^i|/\|v\|, s\right) \\ (l+1)/s & \text{otherwise} \end{cases},$$

and $p(a, s) = as - l$ for any $a \in [0, 1]$. Here, s is the number of quantization levels distributed between 0 and 1, and $l \in [0, s)$ such that $|v^i|/\|v\| \in [l/s, (l+1)/s]$.

Notice that when we let $s = 1$ (and hence $l = 0$) in Definition 4, the low-precision quantizer also reduces to the ternary quantizer defined above.

It is easily shown that these quantizers are sign-preserving unbiased random quantizers. Specifically, we have the following results:

Proposition 1 ([10])**.** *The gradient sparisifier $S : \mathbb{R}^d \to \mathbb{R}^d$ is a sign-preserving URQ, which satisfies*

1. $\mathbf{E}\{\|S(v)\|^2\} \leq (1/p_{\min})\|v\|^2$ *where* $p_{\min} = \min_{i \in [1,d]} p^i$ *, and*
2. $\mathbf{E}\{\|S(v))\|_0\} = \sum_{i=1}^d p^i$.

Proposition 2 (Lemma 3.1 in [11])**.** *The low-precision quantizer $Q_b : \mathbb{R}^d \to \mathbb{R}^d$ is a sign-preserving URQ, which satisfies*

1. $\mathbf{E}\{\|Q_b(v)\|^2\} \leq \left(1 + \min\left(d/s^2, \sqrt{d}/s\right)\right)\|v\|^2$, *and*

2. $\mathbf{E}\{\|Q_b(v))\|_0\} \leq s(s+\sqrt{d})$.

Proposition 1 and 2 both imply that $\mathbf{E}\|Q(v)\|^2$ is close to $\|v\|^2$ if the URQs are sufficiently accurate; *e.g.*, when we set $p^i = 1$ for all i in the gradient sparsifier (we send the full vector) and when we let $s \to \infty$ in the low-precision quantizer (we send the exact solution). Although the probability p^i in the gradient sparsifier can be time-varying (*e.g.*, when we set $p^i \propto v^i$) we assume a time-invariant α-value in the analysis below to simplify notation.

4 Convergence Analysis of Quantized Gradient Method

In this section, we study the impact of gradient compression on the convergence rate guarantees for the gradient descent algorithm. Although this single-master/single-worker architecture is of limited practical interest, it complements and improves on earlier results (e.g. [14]) and establishes a baseline for the distributed master-worker architectures studied later. Explicit formulas for the iteration and communication complexity of GD with URQ compression are also given.

We start by considering the compressed GD algorithm

$$x_{k+1} = x_k - \gamma_k Q(\nabla f(x_k)), \tag{4}$$

where γ_k is a positive step size, and $Q : \mathbb{R}^d \to \mathbb{R}^d$ is a URQ. Throughout this section, we derive explicit expressions for how the variance bound α of the URQ affects admissible step-sizes and guaranteed convergence times. We begin by considering strongly convex optimization problems.

Theorem 1. *Consider the optimization problem* (1) *under Assumption 1, 2 and 3. Suppose that* $\gamma_k = (1/\alpha)\left(2/(\mu+\bar{L})\right)$ *where* $\bar{L} = L\sqrt{m(1+\Delta)}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (4) *satisfy*

$$\mathbf{E}\|x_k - x^\star\|^2 \leq \rho^k \|x_0 - x^\star\|^2,$$

where $\rho = 1 - \frac{1}{\alpha}\frac{4\mu\cdot\bar{L}}{(\mu+\bar{L})^2}$.

Proof. See Appendix B. □

One naive encoding of a vector processed by the URQ requires $c(\log_2 d + B)$ bits: $\log_2 d$ bits to represent each index and B bits to represent the corresponding vector entry of c non-zero values. Hence, Theorem 1 yields the following iteration and communication complexity.

Corollary 1. *Consider the optimization problem* (1) *under Assumption 1, 3 and 2. Suppose that* $\gamma_k = (1/\alpha)\left(2/(\mu+\bar{L})\right)$ *where* $\bar{L} = L\sqrt{m(1+\Delta}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. *Given* $\varepsilon_0 = \|x_0 - x^\star\|^2$, *by running* (4) *for at most*

$$k^\star = \alpha\frac{(\mu+\bar{L})^2}{4\mu\bar{L}}\log\left(\varepsilon_0/\varepsilon\right)$$

iterations, under which

$$B^\star = (\log_2 d + B)\, c\cdot\alpha\frac{(\mu+\bar{L})^2}{4\mu\bar{L}}\log\left(\varepsilon_0/\varepsilon\right)$$

bits are sent, we ensure that $\mathbf{E}\|x_k - x^\star\|^2 \leq \varepsilon$. *Here* B *is the number of bits required to encode a single vector entry and* $\mathbb{E}\{\|Q(v)\|_0\} \leq c$.

Proof. See Appendix C. □

Theorem 1 quantifies how the convergence guarantees depend on α. If the worker node sends the exact gradient, i.e. $Q(\nabla f(x_k)) = \nabla f(x_k)$, $\alpha = 1$ and Theorem 1 recovers the convergence rate result of GD for strongly convex optimization with $\gamma_k = 2/(\mu+\bar{L})$ presented in [18, 19]. If the quantizer produces the less accurate vector (larger α), then we must decrease the step size γ_k to guarantee numerical stability, and accept that the ε-convergence times $T^\star$ will increase. The results are extended to convex optimization problems in Appendix D and E in the supplementary material.

We conclude this section by studying the following compressed IAG algorithm: given an initial point x_0 and a fixed, positive step size γ

$$x_{k+1} = x_k - \gamma Q\left(\sum_{i=1}^{m} \nabla f_i(x_{k-\tau_k^i})\right). \tag{5}$$

The iteration accounts for heterogeneous worker delays, but performs a centralized compression of the sum of staled gradients. We include the result here to highlight how the introduction of heterogeneous delays affect our convergence guarantees, and consider it as an intermediate step towards the more practical architectures studied in the next section. Note that if we let $\tau_k^i = 0$ (and therefore $\tau = 0$), then the compressed IAG iteration (5) reduces to the compressed GD iteration (4).

Theorem 2. *Consider the optimization problem* (1) *under Assumption 1, 3 and 2. Suppose that*

$$\gamma < \min\left(\frac{\mu}{\sqrt{\alpha}\tau\bar{L}^2}, \frac{1}{\alpha\bar{L}}\right),$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$, $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$, *and* $\tau_k^i \leq \tau$ *for all* i, k. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (5) *satisfy*

$$\mathbf{E}[f(x_k) - f(x^\star)] \leq (p+q)^{k/(1+2\tau)}\big(f(x_0) - f(x^\star)\big)$$

where $p = 1 - \mu\gamma$ *and* $q = \bar{L}^4\gamma^3\tau^2\alpha/\mu$.

Proof. See Appendix G. □

The admissible step-sizes in Theorem 2 depends on both the delay bound τ and α. The upper bound on the step-size in Theorem 2 is smaller than the corresponding result in Theorem 1. If the quantizer produces the exact output, then the proposed algorithm coincides with the IAG algorithm (3) for strongly convex optimization. Letting $\alpha = 1$ in our analysis yields a slightly better step size than the one presented in [7]. In this supplementary material (Appendix E), we include a corollary which gives explicit expressions for the associated ε-convergence times and expected information exchange from workers to master.

Furthermore, we extend the result for the optimization problem without the strong convexity assumption as follows:

Theorem 3. *Consider the optimization problem* (1) *under Assumption 1 and 3. Suppose that*

$$\gamma < \frac{1}{\sqrt{1 + 8\left(1 + \beta(1+\theta)\right)\tau(\tau+1)}}\frac{2}{\bar{L}},$$

and $\beta < 1/\left(2(1+1/\theta)\right)$ *for* $\theta > 0$. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (5) *satisfy*

$$\min_{k\in[0,K]} \mathbf{E}\|\nabla f(x_k)\|^2 \leq \frac{1}{a}\frac{1}{K+1}\left(f(x_0) - f^\star\right),$$

where $a = \gamma/2 - \gamma\beta(1+1/\theta)$.

Proof. See Appendix I. □

Theorem 3 implies the sufficient accuracy of the compression techniques to guarantee the numerical stability of the compressed IAG algorithms. Unlike Theorem 2, the step size from this theorem is independent of the conditional number $\bar{L}/\mu$.

5 Distributed Quantized Gradient Method

Before we present the convergence results for the compressed incremental aggregate gradient algorithm, we consider its synchronous counterpart where the master waits for all workers to return before it updates the decision vector. Thus, we study the following algorithm: given the initial point x_0, a positive step size γ_k and the URQ Q, iterates x_k are generated via

$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{m} Q(\nabla f_i(x_k)). \tag{6}$$

Since UQRs are random and modify the gradient vectors and their support, the sparsity patterns of the quantized gradients are time-varying and can be characterized by the quantities

$$\begin{aligned}\Delta_{\max}^k &= \max_{i\in[1,m]}\left\{\sum_{j=1,j\neq i}^{m}\mathbf{1}\left\{\text{supp}(Q(a_i))\cap\text{supp}(Q(a_j))\neq\emptyset\right\}\right\}\\ \Delta_{\text{ave}}^k &= \tfrac{1}{m}\sum_{i=1}^{m}\left\{\sum_{j=1,j\neq i}^{m}\mathbf{1}\{\text{supp}(Q(a_i))\cap\text{supp}(Q(a_j))\neq\emptyset\}\right\}.\end{aligned} \tag{7}$$

A limitation with these quantities is that they cannot be computed off-line. However, since gradient compression reduces the support of vectors, $\text{supp}(Q(a_i))\subset\text{supp}(a_i)$, it always holds that $\Delta_{\max}^k\leq\Delta_{\max}$ and $\Delta_{\text{ave}}^k\leq\Delta_{\text{ave}}$.

The next lemma enables us to benefit from sparsity in our analysis.

Lemma 2. *Under Assumption 3, for $k\geq 0$*

$$\left\|\sum_{i=1}^{m}Q(\nabla f_i(x_k))\right\|^2\leq\sigma_k\sum_{i=1}^{m}\left\|Q(\nabla f_i(x_k))\right\|^2,$$

where

$$\sigma_k=\min\left(\sqrt{m(1+\Delta_{\text{ave}}^k)},1+\Delta_{\max}^k\right).$$

Moreover,

$$\sigma_k\leq\sigma=\min\left(\sqrt{m(1+\Delta_{\text{ave}})},1+\Delta_{\max}\right).$$

Proof. See Appendix F. □

Notice that Lemma 2 quantifies the combined impact of data sparsity and compression. We have $\sigma_k=1$ if the quantized gradients are completely sparse (their support sets do not overlap), whereas $\sigma_k=m$ if the quantized gradients are completely dense (all support sets overlap).

We are now ready to state our convergence result for strongly convex loss functions.

Theorem 4. *Consider the optimization problem* (1)*under Assumption 1, 2 and 3. Suppose that $\gamma=1/\left(L\alpha(1+\theta)\sigma\right)$ for some $\theta>0$. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by* (4) *satisfy*

$$\mathbf{E}\|x_k-x^\star\|^2\leq(1-\mu\gamma)^k\|x_0-x^\star\|^2+\frac{1}{\mu\theta L}\sum_{i=1}^{m}\mathbf{E}\|\nabla f_i(x^\star)\|^2,$$

Proof. See Appendix J. □

Theorem 4 states that the iterates genenrated by D-QGD (6) converge to a ball around the optimal solution. It shows explicitly how the sparsity measure σ and the quantizer accuracy parameter α affect the convergence guarantees. Note that a larger value of θ allows for larger step-sizes and better convergence factor, but also a larger residual error.

For simplicity of notation and applicability of the results, we formulated Theorem 4 in terms of σ and not σ_k (the proof, however, also provides convergence guarantees in terms of σ_k). The result is conservative in the sense that compression increases sparsity of the gradients, which should translate into larger step-sizes. To evaluate the degree of conservatism, we carry out Monte Carlo simulations on the data sets described in Table 2. We indeed note that σ_k is significantly smaller than σ.

6 Q-IAG Method

In this section, we rather consider the quantized version of the optimization algorithm which is suited for communications with limited bandwidth. Therefore, we study the convergence rate of the quantized version of the IAG algorithm (Q-IAG) where the update is

$$x_{k+1}=x_k-\gamma\sum_{i=1}^{m}Q\left(\nabla f_i(x_{k-\tau_k^i})\right), \tag{8}$$

Data Set	σ/m	$\mathbf{E}\{\sigma_k\}/m$ with GS	$\mathbf{E}\{\sigma_k\}/m$ with TQ	$\mathbf{E}\{\sigma_k\}/m$ with LP
RCV1-train	0.83	0.66	0.07	0.42
real-sim	0.8278	0.58	0.06	0.37
GenDense	1	1	0.7	1

Table 1: Empirical evaluations of σ_k and σ with gradient sparsifier (GS) with $p_i = 0.5$, with ternary quantizer (TQ), and with low-precision quantizer (LP) with $s = 4$.

where γ is the constant step size, and Q is the URQ. Notice that

$$\mathbf{E}\left\{\sum_{i=1}^{m} Q\left(\nabla f_i(x_{k-\tau_k^i})\right)\right\} = \sum_{i=1}^{m} \nabla f_i(x_{k-\tau_k^i}).$$

By Assumption 3, $\mathrm{supp}(Q(\nabla f_i(x_{k-\tau_k^i}))) = \mathrm{supp}(Q(a_i))$, and thus the sparsity measures defined (7) will be used to strengthen our main analysis.

Now, we present the result for strongly convex optimization.

Theorem 5. *Consider the optimization problem* (1) *under Assumption 1, 3 and 2. Suppose that*

$$\gamma < \frac{2\mu}{1 + m\sigma\alpha L^2\left(2\bar{L}^2\tau^2 + (1+\theta)\right)},$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$, $\Delta = \min(\Delta_{\mathrm{ave}}, \Delta_{\max})$, $\sigma = \min\left(\sqrt{m(1+\Delta_{\mathrm{ave}})}, 1+\Delta_{\max}\right)$, $\tau_k^i \leq \tau$ *for all* i, k, *and* $\theta > 0$. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (8) *satisfy*

$$\mathbf{E}\|x_k - x^\star\|^2 \leq (p+q)^{k/(1+2\tau)}\|x_0 - x^\star\|^2 + e/(1-p-q),$$

where

$$\begin{aligned} p &= 1 - 2\mu\gamma + \gamma^2 \\ q &= 2m\sigma\alpha L^2\gamma^2\bar{L}^2\tau^2 + (1+\theta)\gamma^2 m\alpha\sigma L^2 \\ e &= \left(2m\alpha\gamma^2\bar{L}^2\tau^2 + (1+1/\theta)\gamma^2\sigma\alpha\right)\sum_{i=1}^{m}\mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2. \end{aligned}$$

Proof. See Appendix L. □

Unlike the result for the compressed IAG algorithm (5), Theorem (5) can only guarantee that the Q-IAG algorithm (8) converges to within a ball around the optimum. In particular, the step-size shows the dependency of the delay τ and of the parameter related to the choice of URQ through the β parameter. In the absence of the worker asynchrony ($\tau = 0$), the upper bound on the step-size becomes $2\mu/((1+\theta)m\sigma\alpha L^2)$, which is smaller than the step size allowed by Theorem 4.

Next, we present the result for optimization problems without the strong convexity assumption of the objective function f. However, we need to impose the bounded gradient assumption for analyze the Q-IAG algorithms under this problem setting.

Assumption 4. *There exists a scalar* C *such that*

$$\|\nabla f_i(x)\| \leq C,$$

for any component function $f_i : \mathbb{R}^d \to \mathbb{R}$ *and* $x \in \mathbb{R}^d$.

Now, the result is shown below:

Theorem 6. *Consider the optimization problem* (1) *under Assumption 1, 3 and 4. Suppose that*

$$\gamma < \frac{1}{1+\sqrt{1+8\tau(\tau+1)}}\frac{2}{\bar{L}},$$

and $\bar{L} = L\sqrt{m(1+\Delta)}$, $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$, *and* $\tau_k^i \leq \tau$ *for all* i, k. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (8) *satisfy*

$$\min_{k\in[0,K]} \mathbf{E}\|\nabla f(x_k)\|^2 \leq \frac{2}{\gamma}\frac{1}{K+1}\left(f(x_0) - f^\star\right) + e,$$

where $e = 2\beta\sigma m C^2$.

Proof. See Appendix M.

□

Unlike Theorem 5, the step size stated in Theorem 6 does not depend on the conditional number $\bar{L}/\mu$.

7 Simulation Results

We consider the empirical risk minimization problem (1) with component loss functions on the form of

$$f_i(x) = \frac{1}{2\rho}\|A_i x - b_i\|^2 + \frac{\sigma}{2}\|x\|^2,$$

where $A_i \in \mathbb{R}^{p\times d}$ and $b_i \in \mathbb{R}^p$. We distributed data samples $(a_1, b_1), \ldots, (a_n, b_n)$ among m workers. Hence, $n = mp$. The experiments were done using both synthetic and real-world data sets as shown in Table 2. Each data sample a_i is then normalized by its own Euclidean norm. We evaluated the performance of the distributed gradient algorithms (4)-(8) using the gradient sparsifier, the low-precision quantizer and the ternary quantizer in Julia. We set $m = 3$, $x_0 = \mathbf{0}$, set $\sigma = 1$, and set ρ equal to the total number of data samples according to Table 2. In addition, GenDense from Table 2 generated the dense data set such that each element of the data matrices A_i is randomly drawn from a uniform random number between 0 and 1, and each element of the class label vectors b_i is the sign of a zero-mean Gaussian random number with unit variance. For the gradient sparsifier, we assumed that vector elements are represented by 64 bits (IEEE doubles) while the low-precision quantizer only requires $1 + \log_2(s)$ bits to encode each vector entry. For the distributed algorithms, we have used $\tau = m$.

Data Set	Type	Samples	Dimension
RCV1-train	sparse	23149	47236
real-sim	sparse	72309	20958
covtype	dense	581012	54
GenDense	dense	40000	1000

Table 2: Summary of synthetic and real-world data sets used in our experiments.

Figure 1 and 2 show the trade-off between the convergence in terms of iteration count and the number of communicated bits. Naturally, the full gradient method has the fastest convergence, and the ternary quantizer is slowest. The situation is reversed if we judge the convergence relative to the number of communicated bits. In this case, the ternary quantizer makes the fastest progress per information bit, followed by the 3-bit low-precision quantizer ($s = 4$). In fact, the full gradient descent requires more bits in the order of magnitude to make 50% progress than the ternary quantizer.

The corresponding results for Q-IAG in the asynchronous parameter server setting are shown in Figure 3 and 4. The results are qualitatively similar: sending the gradient vectors in higher precision yields the fastest convergence but can be extremely wasteful in terms of communication load. The low-precision quantizer allows us to make a gentle trade-off between the two objectives, having both a rapid and communication-efficient convergence. In particular, the results from covtype show that a fast convergence in terms of both iteration counts and communications load for the low-precision quantizer with the higher number of quantization levels.

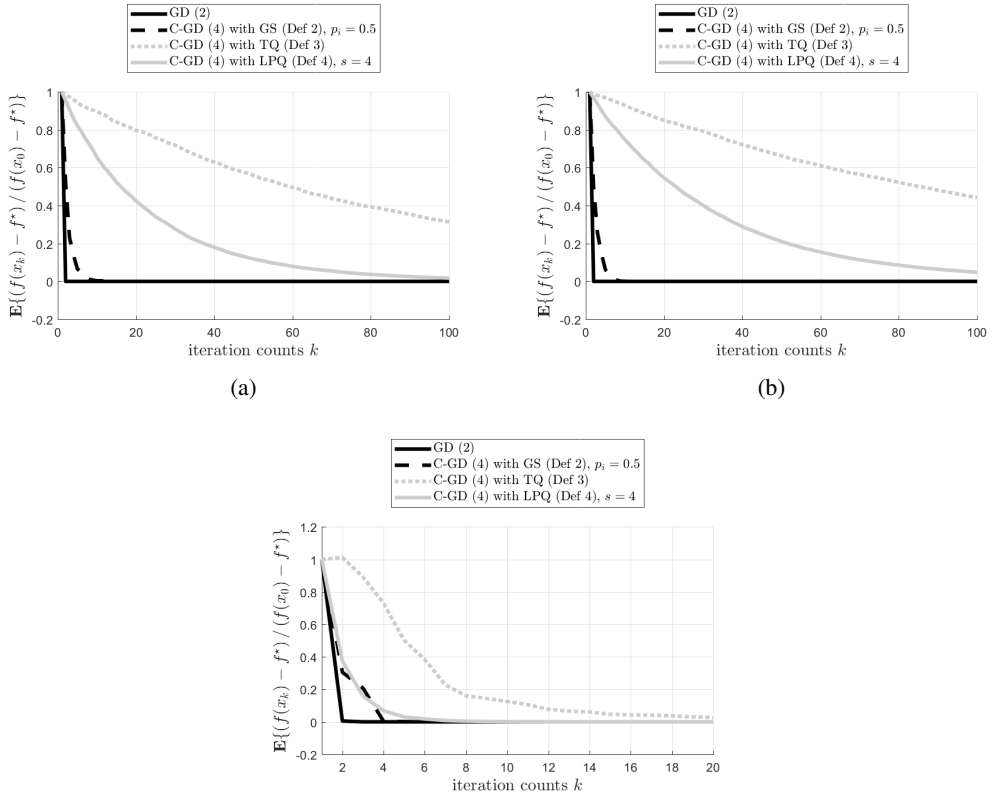

Figure 1: Convergence of compressed gradient descent algorithms (4) using different compression techniques over real-world data sets; that is, (a) real-sim, (b) RCV1-train and (c) covtype.

8 Conclusions and Future Work

We have established a unified framework for both synchronous and asynchronous distributed optimization using compressed gradients. The framework builds on the concept of unbiased randomized quantizers (URQs), a class of gradient compression schemes which cover several important proposals from the literature [10, 11]. We have established non-asymptotic convergence rate guarantees for both GD and IAG with URQ compression. The convergence rate guarantees give explicit formulas for how quantization accuracy and staleness bounds affect the expected time to reach an ε-optimal solution. These results allowed us to characterize the trade-off between iteration and communication complexity of gradient descent under gradient compression.

We are currently working on extending the framework to allow for deterministic quantizers. Such quantizers are not necessarily unbiased, but satisfy additional inequalities which could be useful for the analysis. Another research direction is to establish non-asymptotic convergence rates under quantization-error compensation, which have been reported to work well in empirical studies [17]. Finally, we would also like to analyze the effect of compressing the traffic from master to workers.

9 Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

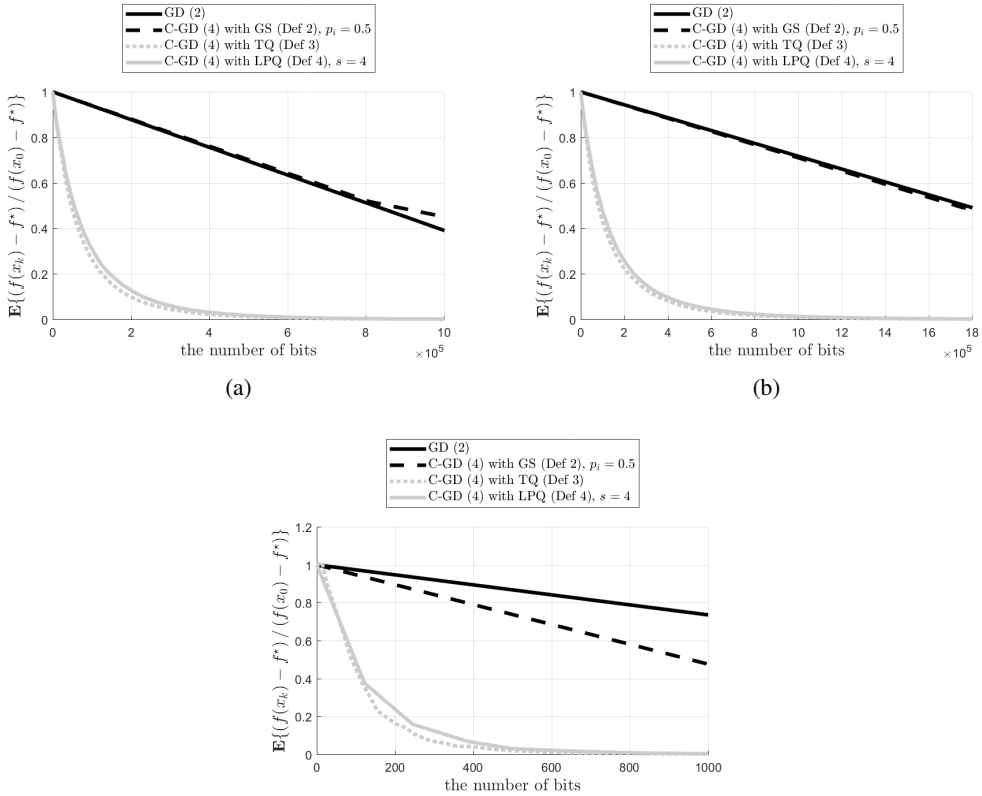

Figure 2: Convergence of compressed gradient descent algorithms (4) using different compression techniques over real-world data sets; that is, (a) real-sim, (b) RCV1-train and (c) covtype.

References

[1] Shamir, Ohad. "Without-replacement sampling for stochastic gradient methods." In *Advances in Neural Information Processing Systems*, pp. 46-54. 2016.

[2] Needell, Deanna, Rachel Ward, and Nati Srebro. "Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm." In *Advances in Neural Information Processing Systems*, pp. 1017-1025. 2014.

[3] Zhang, Ruiliang, and James Kwok. "Asynchronous distributed ADMM for consensus optimization." In *International Conference on Machine Learning*, pp. 1701-1709. 2014.

[4] Li, Mu, David G. Andersen, Alexander J. Smola, and Kai Yu. "Communication efficient distributed machine learning with the parameter server." In *Advances in Neural Information Processing Systems*, pp. 19-27. 2014.

[5] Chen, Jianmin, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. "Revisiting distributed synchronous SGD." *arXiv preprint arXiv:1604.00981*, 2016.

[6] Blatt, Doron, Alfred O. Hero, and Hillel Gauchman. "A convergent incremental gradient method with a constant step size." *SIAM Journal on Optimization* 18, no. 1 (2007): 29-51.

[7] Gurbuzbalaban, Mert, Asuman Ozdaglar, and Pablo A. Parrilo. "On the convergence rate of incremental aggregated gradient algorithms." *SIAM Journal on Optimization* 27, no. 2 (2017): 1035-1048.

[8] Tseng, Paul, and Sangwoon Yun. "Incrementally updated gradient methods for constrained and regularized optimization." *Journal of Optimization Theory and Applications* 160, no. 3 (2014): 832-853.

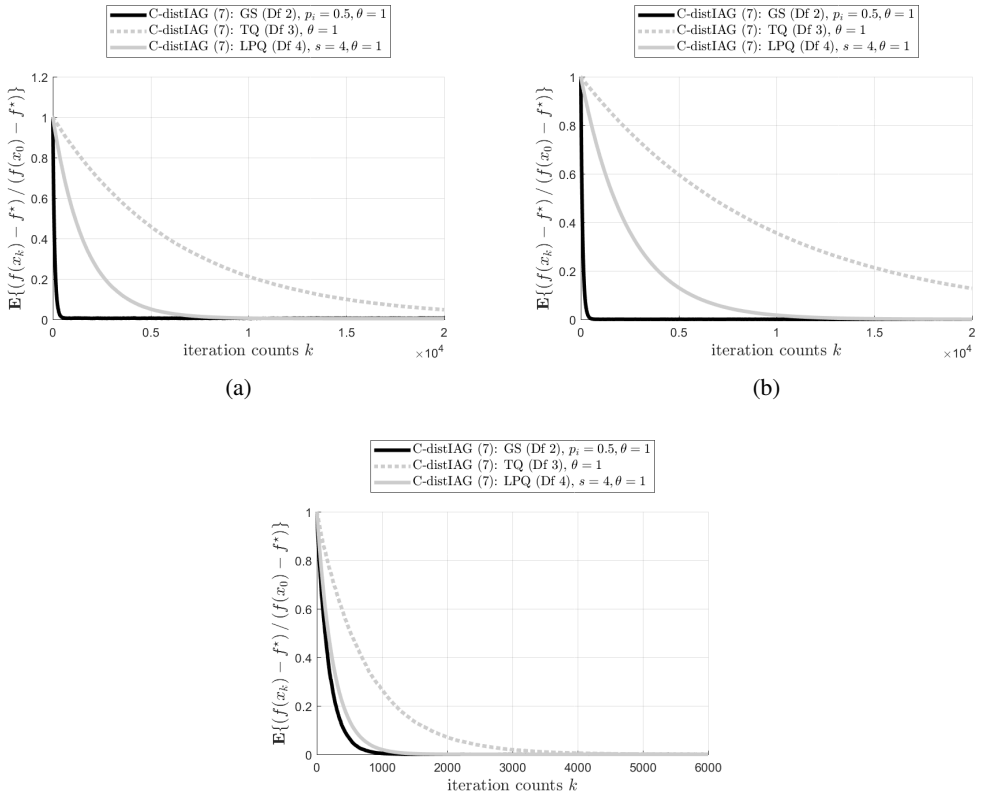

Figure 3: Convergence of Q-IAG algorithms (8) using different compression techniques over real-world data sets; that is, (a) real-sim, (b) RCV1-train and (c) covtype.

[9] Aytekin, Arda, Hamid Reza Feyzmahdavian, and Mikael Johansson. "Analysis and implementation of an asynchronous optimization algorithm for the parameter server." *arXiv preprint arXiv:1610.05507*, 2016.

[10] Wangni, Jianqiao, Jialei Wang, Ji Liu, and Tong Zhang. "Gradient Sparsification for Communication-Efficient Distributed Optimization." *arXiv preprint arXiv:1710.09854*, 2017.

[11] Alistarh, Dan, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding." In *Advances in Neural Information Processing Systems*, pp. 1707-1718. 2017.

[12] De Sa, Christopher M., Ce Zhang, Kunle Olukotun, and Christopher Ré. "Taming the wild: A unified analysis of hogwild-style algorithms." In *Advances in neural information processing systems*, pp. 2674-2682. 2015.

[13] Wen, Wei, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. "Terngrad: Ternary gradients to reduce communication in distributed deep learning." In *Advances in Neural Information Processing Systems*, pp. 1508-1518. 2017.

[14] Magnússon, Sindri, Chinwendu Enyioha, Na Li, Carlo Fischione, and Vahid Tarokh. "Convergence of limited communications gradient methods." *IEEE Transactions on Automatic Control*, 2017.

[15] Nguyen, Lam M., Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. "SGD and Hogwild! Convergence Without the Bounded Gradients Assumption." *arXiv preprint arXiv:1802.03801*, 2018.

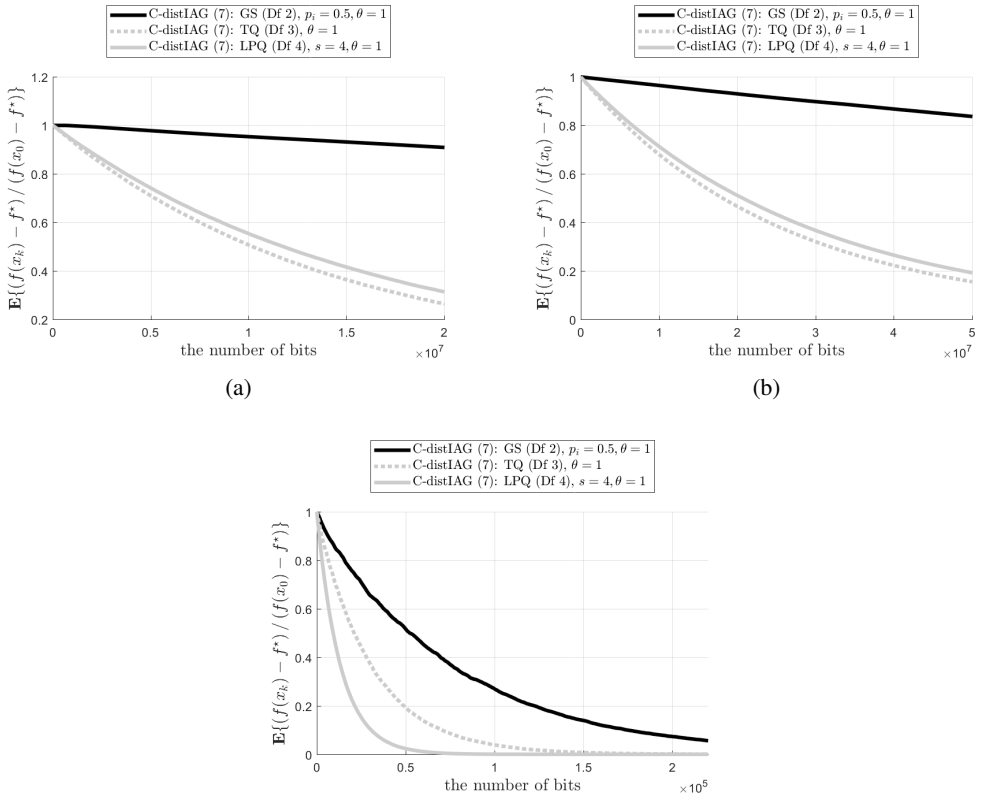

Figure 4: Convergence of Q-IAG algorithms (4) using different compression techniques over real-world data sets; that is, (a) real-sim, (b) RCV1-train and (c) covtype.

[16] Tsitsiklis, John N., and Zhi-Quan Luo. "Communication complexity of convex optimization." *Journal of Complexity* 3, no. 3 (1987): 231-243.

[17] Seide, Frank, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs." In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[18] Polyak, Boris T. "Introduction to optimization. Translations series in mathematics and engineering." *Optimization Software*, 1987.

[19] Nesterov, Yurii. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

[20] Khirirat, Sarit, Hamid Reza Feyzmahdavian, and Mikael Johansson. "Mini-batch gradient descent: Faster convergence under data sparsity." In *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*, pp. 2880-2887. IEEE, 2017.

[21] Aytekin, Arda, Hamid Reza Feyzmahdavian, and Mikael Johansson. "Asynchronous incremental block-coordinate descent." In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pp. 19-24. IEEE, 2014.

A Proof of Lemma 1

For $x, y \in \mathbb{R}^d$, $\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle$. This property implies that

$$\begin{aligned}
&\|\nabla f(x) - \nabla f(y)\|^2 \\
&= \left\| \sum_{i=1}^m \nabla f_i(x) - \sum_{i=1}^m \nabla f_i(y) \right\|^2 \\
&= \|\nabla f_1(x) - \nabla f_1(y)\|^2 + \sum_{i=2}^m \langle \nabla f_1(x) - \nabla f_1(y), \nabla f_i(x) - \nabla f_i(y)\rangle \\
&\quad + \|\nabla f_2(x) - \nabla f_2(y)\|^2 + \sum_{i=1, i\neq 2}^m \langle \nabla f_2(x) - \nabla f_2(y), \nabla f_i(x) - \nabla f_i(y)\rangle + \ldots \\
&\quad + \|\nabla f_m(x) - \nabla f_m(y)\|^2 + \sum_{i=1}^{m-1} \langle \nabla f_m(x) - \nabla f_m(y), \nabla f_i(x) - \nabla f_i(y)\rangle \\
&= \sum_{i=1}^m \|\nabla f_i(x) - \nabla f_i(y)\|^2 + \sum_{i=1}^m \sum_{j=1, j\neq i}^m \langle \nabla f_i(x) - \nabla f_i(y), \nabla f_j(x) - \nabla f_j(y)\rangle e_{i,j},
\end{aligned}$$

where $e_{i,j} = 1$ if $\text{supp}(\nabla f_i(x)) \cap \text{supp}(\nabla f_j(x)) \neq \emptyset$ and 0 otherwise. By Cauchy-Schwarz's inequality, we have

$$\begin{aligned}
&\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \sum_{i=1}^m \|\nabla f_i(x) - \nabla f_i(y)\|^2 + \sum_{i=1}^m \sum_{j=1, j\neq i}^m \|\nabla f_i(x) - \nabla f_i(y)\| \cdot \|\nabla f_j(x) - \nabla f_j(y)\| e_{i,j} \\
&\leq \sum_{i=1}^m L^2 \|x-y\|^2 + \sum_{i=1}^m \sum_{j=1, j\neq i}^m L^2 \|x-y\| \cdot \|x-y\| e_{i,j} \\
&= \left(\sum_{i=1}^m L^2 + \sum_{i=1}^m \sum_{j=1, j\neq i}^m L^2 e_{i,j} \right) \|x-y\|^2 \\
&= L^2 \left(m + \sum_{i=1}^m \sum_{j=1, j\neq i}^m e_{i,j} \right) \|x-y\|^2 \\
&= L^2 m \left(1 + \Delta_{\text{ave}}\right) \|x-y\|^2,
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of the gradients of component functions f_i. Notice that the sparsity pattern of $\nabla f_i(x)$ can be found using the data matrix A, [20].

Next, we can tighten the bound using the maximum conflict degree $\Delta_{\max}$. By Cauchy-Schwarz's inequality and by the Lipschitz gradient assumption of f_i, we have

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\|^2 &\leq L^2 \left(m + \sum_{i=1}^m \sum_{j=1, j\neq i}^m e_{i,j} \right) \|x-y\|^2 \\
&\leq L^2 m \left(1 + \Delta_{\max}\right) \|x-y\|^2,
\end{aligned}$$

where the last inequality derives from the definition of the maximum conflict graph degree. In conclusion,

$$\bar{L}^2 = L^2 m \left(1 + \Delta\right),$$

where $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$.

B Proof of Theorem 1

Using the distance between the iterates $\{x_k\}_{k\in\mathbb{N}}$ and the optimum $x^\star$, we have

$$\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 - 2\gamma_k \langle Q(\nabla f(x_k)), x_k - x^\star\rangle + \gamma_k^2 \|Q(\nabla f(x_k))\|^2.$$

Taking the expectation with respect to all the randomness in the algorithm yields

$$\begin{aligned}
\mathbf{E}\|x_{k+1} - x^\star\|^2 &= \mathbf{E}\|x_k - x^\star\|^2 - 2\gamma_k \mathbf{E}\langle \nabla f(x_k), x_k - x^\star\rangle + \gamma_k^2 \mathbf{E}\|Q(\nabla f(x_k))\|^2 \\
&\leq \mathbf{E}\|x_k - x^\star\|^2 - 2\gamma_k \mathbf{E}\langle \nabla f(x_k), x_k - x^\star\rangle + \gamma_k^2 \alpha \mathbf{E}\|\nabla f(x_k)\|^2,
\end{aligned}$$

where the inequality follows from the second property in Definition 1. It follows from [19, Theorem 2.1.12] that

$$\begin{aligned}\mathbf{E}\|x_{k+1}-x^\star\|^2 &\leq \left(1-2\gamma_k\frac{\mu\bar{L}}{\mu+\bar{L}}\right)\mathbf{E}\|x_k-x^\star\|^2+\left(-2\gamma_k\frac{1}{\mu+\bar{L}}+\gamma_k^2\alpha\right)\mathbf{E}\|\nabla f(x_k)\|^2\\ &=\rho_k\mathbf{E}\|x_k-x^\star\|^2+\left(-2\gamma_k\frac{1}{\mu+\bar{L}}+\gamma_k^2\alpha\right)\mathbf{E}\|\nabla f(x_k)\|^2,\end{aligned}$$

where $\rho_k=1-2\gamma_k\frac{\mu\bar{L}}{\mu+\bar{L}}$. If $-2\gamma_k/(\mu+\bar{L})+\gamma_k^2\alpha\leq 0$, or equivalently

$$\gamma_k\in\left[0,\frac{2}{\alpha(\mu+\bar{L})}\right],$$

then $\rho_k\in[0,1)$ for $\alpha\geq 1$, and the second term on the right-hand side of the above inequality is non-positive. Therefore,

$$\mathbf{E}\|x_{k+1}-x^\star\|^2\leq\rho_k\mathbf{E}\|x_k-x^\star\|^2.$$

This implies that

$$\mathbf{E}\|x_k-x^\star\|^2\leq\left(\prod_{i=1}^k\rho_i\right)\|x_0-x^\star\|^2,\quad k\in\mathbb{N}.$$

If the step-sizes are constant ($\gamma_k=\gamma$ for all k), then

$$\mathbf{E}\|x_k-x^\star\|^2\leq\rho^k\|x_0-x^\star\|^2,\quad k\in\mathbb{N}.$$

C Proof of Corollary 1

From Theorem 1, we get $V_k\leq\rho^k\varepsilon_0$ with $V_k=\mathbf{E}\|x_k-x^\star\|^2$, or equivalently

$$\left(1-\frac{1}{\alpha}\frac{4\mu\cdot\bar{L}}{(\mu+\bar{L})^2}\right)^k\varepsilon_0\leq\varepsilon.$$

Since $-1/\log(1-x)\leq 1/x$ for $0<x\leq 1$ and $\rho\in(0,1)$, we reach the upper bound of $k^\star$. In addition, assume that the number of non-zero elements is at most c. Therefore, the number of bits required to code the vector is at most $(\log_2 d+B)\,c$ bits in each iteration, where B is the number bits required to encode a single vector entry. Hence, we reach the upper bound of $B^\star$.

D Convergence Result of Compressed GD Algorithm for Convex Optimization

Theorem 7. *Consider the optimization problem* (1) *under Assumption 1 and 3. Suppose that* $\gamma_k=(1/\bar{L}\alpha)$ *where* $\bar{L}=L\sqrt{m(1+\Delta)}$ *and* $\Delta=\min(\Delta_{\text{ave}},\Delta_{\max})$. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (4) *satisfy*

$$\mathbf{E}\left(f(x_T)-f^\star\right)\leq\frac{\alpha\bar{L}}{2(T+1)}\|x_0-x^\star\|^2.$$

Proof. Following the proof in Theorem 1, we reach

$$\mathbf{E}\|x_{k+1}-x^\star\|^2\leq\mathbf{E}\|x_k-x^\star\|^2-2\gamma_k\mathbf{E}\langle\nabla f(x_k),x_k-x^\star\rangle+\gamma_k^2\alpha\mathbf{E}\|\nabla f(x_k)\|^2.$$

By the property of Lipschitz continuity of $\nabla f(x)$, we have:

$$\langle\nabla f(x_k),x_k-x^\star\rangle\geq f(x_k)-f^\star+\frac{1}{2\bar{L}}\|\nabla f(x_k)\|^2.$$

This implies that

$$\mathbf{E}\|x_{k+1}-x^\star\|^2\leq\mathbf{E}\|x_k-x^\star\|^2-2\gamma_k\mathbf{E}\left(f(x_k)-f^\star\right)+\left(-\frac{\gamma_k}{\bar{L}}+\gamma_k^2\alpha\right)\mathbf{E}\|\nabla f(x_k)\|^2.$$

Next, assume that $-\gamma_k/\bar{L} + \gamma_k^2\alpha \leq 0$ or equivalently $\gamma_k \leq 1/(\bar{L}\alpha)$. Due to the non-negativity of the Euclidean norm,

$$\mathbf{E}\|x_{k+1} - x^\star\|^2 \leq \mathbf{E}\|x_k - x^\star\|^2 - 2\gamma_k \mathbf{E}\left(f(x_k) - f^\star\right).$$

After the manipulation, we have:

$$2\sum_{k=0}^{T} \gamma_k \mathbf{E}\left(f(x_k) - f^\star\right) \leq \mathbf{E}\|x_0 - x^\star\|^2 - \mathbf{E}\|x_{T+1} - x^\star\|^2. \tag{9}$$

Again from the Lipschitz gradient assumption of f, we have

$$f(x_{k+1}) \leq f(x_k) - \gamma_k \langle \nabla f(x_k), Q(\nabla f(x_k))\rangle + \frac{\bar{L}\gamma_k^2}{2}\|Q(\nabla f(x_k))\|^2.$$

Taking the expectation over all random variables yields

$$\mathbf{E}f(x_{k+1}) \leq \mathbf{E}f(x_k) - \left(\gamma_k - \frac{\bar{L}\alpha\gamma_k^2}{2}\right)\|\nabla f(x_k)\|^2,$$

where we reach the inequality by properties stated in Definition 1. Due to the fact that $\gamma_k \leq 1/(\bar{L}\alpha)$ and the non-negativity of the Euclidean norm, we can conclude that $\mathbf{E}f(x_{k+1}) \leq \mathbf{E}f(x_k)$. From (9),

$$2\gamma_{\min}(T+1)\mathbf{E}\left(f(x_T) - f^\star\right) \leq \mathbf{E}\|x_0 - x^\star\|^2 - \mathbf{E}\|x_{T+1} - x^\star\|^2,$$

or equivalently

$$\mathbf{E}\left(f(x_T) - f^\star\right) \leq \frac{1}{2\gamma_{\min}(T+1)}\mathbf{E}\|x_0 - x^\star\|^2,$$

where $\gamma_{\min} = \min\limits_{k\in[0,T]} \gamma_k$. Plugging $\gamma_{\min} = 1/(\bar{L}\alpha)$ yields the result. □

E Complexity of Compressed GD Algorithm for Convex Optimization

Corollary 2. *Consider the optimization problem* (1) *under Assumption 1 and 3. Suppose that* $\gamma_k = (1/\bar{L}\alpha)$ *where* $\bar{L} = L\sqrt{m(1+\Delta)}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. *Given* $\varepsilon_0 = \|x_0 - x^\star\|^2$, *by running* (4) *for at most*

$$T^\star = \frac{\alpha\bar{L}}{2} \cdot \frac{\varepsilon_0}{\varepsilon}$$

iterations, under which

$$B^\star = \left(\log_2 d + B\right) c \cdot \frac{\alpha\bar{L}}{2} \cdot \frac{\varepsilon_0}{\varepsilon}$$

bits are sent, we ensure $\mathbf{E}\left(f(x_T) - f^\star\right) \leq \varepsilon$. *Here* B *is the number of bits required to encode a single vector entry and* $\mathbb{E}\{\|Q(v)\|_0\} \leq c$.

Proof. The upper bound of $T^\star$ is easily obtained by using the inequality in Theorem 7. Also, assume that the number of non-zero elements is at most c. Therefore, the number of bits required to code the vector is at most $\left(\log_2 d + B\right) c$ bits in each iteration, where B is the number bits required to encode a single vector entry. Hence, we reach the upper bound of $B^\star$. □

F Proof of Lemma 2

Denote $s_i^k = \text{supp}(Q(a_i))$. By Assumption 3 and the definition of the Euclidean norm,

$$\begin{aligned}
\left\|\sum_{i=1}^{m} Q(\nabla f_i(x_k))\right\|^2 &= \sum_{i=1}^{m}\|Q(\nabla f_i(x_k))\|^2 + \sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\langle Q(\nabla f_i(x_k)), Q(\nabla f_j(x_k))\rangle \\
&\leq \sum_{i=1}^{m}\|Q(\nabla f_i(x_k))\|^2 \\
&\quad + \underbrace{\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\|Q(\nabla f_i(x_k))\|\|Q(\nabla f_j(x_k))\|\mathbf{1}\left(s_i^k \cap s_j^k \neq \emptyset\right)}_{T}
\end{aligned}$$

where we reach the inequality by Cauchy-Schwarz's inequality. For simplicity, let $e_{i,j} = \mathbf{1}\left(s_i^k \cap s_j^k \neq \emptyset\right)$. Now, we bound the left-hand side using two different data sparsity measures. First, we bound T using the maximum conflict degree $\Delta_{\max}^k = \max_{i\in[1,m]} \Delta_i$ where $\Delta_i = \sum_{j=1,j\neq i}^m e_{i,j}$. By the fact that $2ab \leq a^2 + b^2$ for $a, b \in \mathbb{R}$, we have

$$\begin{aligned} T &\leq \frac{1}{2}\sum_{i=1}^m \sum_{j=1,j\neq i}^m \left(\|Q(\nabla f_i(x_k))\|^2 + \|Q(\nabla f_j(x_k))\|^2\right) e_{i,j} \\ &\leq \Delta_{\max}^k \sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2. \end{aligned}$$

Therefore,

$$\left\|\sum_{i=1}^m Q(\nabla f_i(x_k))\right\|^2 \leq (1+\Delta_{\max}^k)\sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2.$$

Next, we bound the left-hand side using the average conflict degree $\Delta_{\text{ave}}^k = (1/m)\sum_{i=1}^m \Delta_i$. By Cauchy-Schwarz's inequality, we get:

$$\begin{aligned} \left\|\sum_{i=1}^m Q(\nabla f_i(x_k))\right\|^2 &= \sum_{i=1}^m \|Q(\nabla f_i(x_k)\| \sum_{j=1}^m \|Q(\nabla f_j(x_k))\| e_{i,j} \\ &\leq \sum_{i=1}^m \|Q(\nabla f_i(x_k))\| \sqrt{\sum_{j=1}^m \|Q(\nabla f_j(x_k))\|^2 \sum_{j=1}^m e_{i,j}^2} \\ &\leq \sqrt{\sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2} \sqrt{\sum_{j=1}^m \|Q(\nabla f_j(x_k))\|^2} \sqrt{\sum_{i=1}^m \sum_{j=1}^m e_{i,j}^2} \\ &\leq \sqrt{\sum_{i=1}^m \sum_{j=1}^m e_{i,j}} \sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2 \\ &= \sqrt{m(1+\Delta_{\text{ave}}^k)} \sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2. \end{aligned}$$

In conclusion,

$$\left\|\sum_{i=1}^m Q(\nabla f_i(x_k))\right\|^2 \leq \sigma_k \sum_{i=1}^m \|Q(\nabla f_i(x_k))\|^2,$$

where $\sigma_k = \min\left(\sqrt{m(1+\Delta_{\text{ave}}^k)}, 1+\Delta_{\max}^k\right)$.

G Proof of Theorem 2

Denote $g_k = \sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i})$. Let us first introduce two main lemmas which are instrumental to our main analysis.

Lemma 3. *Consider the iterates generated by* (5)*. For* $k \in \mathbb{N}_0$*,*

$$\|g_k\|^2 \leq \frac{2\bar{L}^2}{\mu} \max_{s\in[k-\tau,k]} f(x_s) - f(x^\star),$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$.

Proof. Since $\nabla f(x^\star) = 0$, we have

$$\|g_k\|^2 = \left\|\sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i}) - \nabla f_i(x^\star)\right\|^2.$$

Following the proof of Lemma 1 with $x = x_{k-\tau_k^i}$ and $y = x^\star$ yields

$$\|g_k\|^2 \leq \bar{L}^2 \max_{s\in[k-\tau,k]} \|x_s - x^\star\|^2,$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$ and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. The result also uses the fact that

$$\|x_{k-\tau_k^i} - x^\star\| \leq \max_{s\in[k-\tau,k]} \|x_s - x^\star\|.$$

Since

$$f(x) - f(x^\star) \geq \frac{\mu}{2}\|x - x^\star\|^2,$$

for any x, it follows that

$$\|g_k\|^2 \leq \frac{2\bar{L}^2}{\mu} \max_{s\in[k-\tau,k]} f(x_s) - f(x^\star).$$

□

Lemma 4. *The sequence $\{x_k\}$ generated by* (5) *satisfies*

$$\mathbf{E}\|\nabla f(x_k) - g_k\|^2 \leq \frac{2\gamma^2\bar{L}^4\tau^2\alpha}{\mu} \max_{s\in[k-2\tau,k]} f(x_s) - f(x^\star),$$

for $k \in \mathbb{N}_0$, where $\bar{L} = L\sqrt{m(1+\Delta)}$ and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$.

Proof. By the definition of g_k,

$$\|\nabla f(x_k) - g_k\|^2 = \left\| \sum_{i=1}^m \nabla f_i(x_k) - \nabla f_i(x_{k-\tau_k^i}) \right\|^2.$$

Following the proof of Lemma 1 with $x = x_k$ and $y = x_{k-\tau_k^i}$ yields

$$\|\nabla f(x_k) - g_k\|^2 \leq \bar{L}^2 \max_{i\in[1,m]} \left\| x_k - x_{k-\tau_k^i} \right\|^2,$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$ and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. We also reach the result by the fact that

$$\left\| x_k - x_{k-\tau_k^i} \right\| \leq \max_{i\in[1,m]} \left\| x_k - x_{k-\tau_k^i} \right\|.$$

Next, notice that

$$\begin{aligned}
\|\nabla f(x_k) - g_k\|^2 &\leq \bar{L}^2 \max_{i\in[1,m]} \left\| \sum_{j=k-\tau_k^i}^{k-1} x_{j+1} - x_j \right\|^2 \\
&\leq \bar{L}^2 \max_{i\in[1,m]} \tau_k^i \sum_{j=k-\tau_k^i}^{k-1} \|x_{j+1} - x_j\|^2 \\
&\leq \bar{L}^2\tau \sum_{j=k-\tau}^{k-1} \|x_{j+1} - x_j\|^2 \\
&= \bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \|Q(g_j)\|^2.
\end{aligned}$$

The second inequality derives from the bounded delay assumption. Taking the expectation with respect to the randomness yields

$$\mathbf{E}\|\nabla f(x_k) - g_k\|^2 \leq \gamma^2\bar{L}^2\tau\alpha \sum_{j=k-\tau}^{k-1} \|g_j\|^2.$$

It follows from Lemma 3 that

$$\begin{aligned}\mathbf{E}\|\nabla f(x_k) - g_k\|^2 &\leq \frac{2\gamma^2\bar{L}^4\tau\alpha}{\mu}\sum_{j=k-\tau}^{k-1}\max_{s\in[j-\tau,j]} f(x_s) - f(x^\star) \\ &\leq \frac{2\gamma^2\bar{L}^4\tau^2\alpha}{\mu}\max_{s\in[k-2\tau,k]} f(x_s) - f(x^\star).\end{aligned}$$

□

We now prove Theorem 2. Since the entire cost function f has Lipschitz continuous gradient with constant $\bar{L}$, we have

$$f(x_{k+1} - f(x^\star) \leq f(x_k) - f(x^\star) - \gamma\langle Q(g_k), \nabla f(x_k)\rangle + \frac{\gamma^2\bar{L}}{2}\|Q(g_k)\|^2.$$

Taking the expectation with respect to the randomness and using the second property in Definition 1, we obtain

$$\mathbf{E}[f(x_{k+1} - f(x^\star)] \leq \mathbf{E}[f(x_k) - f(x^\star)] - \gamma\mathbf{E}[\langle g_k, \nabla f(x_k)\rangle] + \frac{\gamma^2\alpha\bar{L}}{2}\mathbf{E}[\|g_k\|^2].$$

If $\gamma\alpha\bar{L} \leq 1$, then $\gamma^2\alpha\bar{L} \leq \gamma$, which implies that

$$\mathbf{E}[f(x_{k+1} - f(x^\star)] \leq \mathbf{E}[f(x_k) - f(x^\star)] - \gamma\mathbf{E}[\langle g_k, \nabla f(x_k)\rangle] + \frac{\gamma}{2}\mathbf{E}[\|g_k\|^2].$$

Using $g_k = g_k - \nabla f(x_k) + \nabla f(x_k)$, we have

$$\begin{aligned}\mathbf{E}[f(x_{k+1} - f(x^\star)] &\leq \mathbf{E}[f(x_k) - f(x^\star)] - \frac{\gamma}{2}\mathbf{E}[\|\nabla f(x_k)\|^2] + \frac{\gamma}{2}\mathbf{E}[\|g_k - \nabla f(x_k)\|^2] \\ &\leq (1-\gamma\mu)\mathbf{E}[f(x_k) - f(x^\star)] + \frac{\gamma}{2}\mathbf{E}[\|g_k - \nabla f(x_k)\|^2],\end{aligned}$$

where the second inequality follows from the fact that

$$f(x) - f(x^\star) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2,$$

for any x. It follows from Lemma 4 that

$$\mathbf{E}[f(x_{k+1} - f(x^\star)] \leq (1-\gamma\mu)\mathbf{E}[f(x_k) - f(x^\star)] + \frac{\gamma^3\bar{L}^4\tau^2\alpha}{\mu}\max_{s\in[k-2\tau,k]} f(x_s) - f(x^\star).$$

This inequality can be rewritten as

$$V_{k+1} \leq pV_k + q\max_{s\in[k-2\tau,k]} V_s,$$

where

$$\begin{aligned}V_k &= \mathbf{E}[f(x_k) - f(x^\star)] \\ p &= 1 - \gamma\mu \\ q &= \frac{\gamma^3\bar{L}^4\tau^2\alpha}{\mu}.\end{aligned}$$

According to Lemma 1 of [21], if $p + q < 1$, or, equivalently,

$$\gamma < \frac{\mu}{\bar{L}^2\tau\sqrt{\alpha}},$$

then $V_k \leq (p+q)^{k/(1+2\tau)}V_0$. This completes the proof.

H Complexity of Compressed IAG Algorithm for Strongly Convex Optimization

Corollary 3. *Consider the optimization problem* (1) *under Assumption 1, 3 and 2. Suppose that* $\gamma < \min\left(\mu/(\sqrt{\alpha}\tau\bar{L}^2), 1/(\alpha\bar{L})\right)$, *where* $\bar{L} = L\sqrt{m(1+\Delta)}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. *Given* $\varepsilon_0 = f(x_0) - f^\star$, *by running* (5) *for at most*

$$k^\star = (1+2\tau)\frac{\mu}{\gamma\left(\mu^2 - \bar{L}^4\gamma^2\tau^2\alpha\right)}\log(\varepsilon_0/\varepsilon)$$

iterations, under which

$$B^\star = (\log_2 d + B)c \cdot (1+2\tau)\frac{\mu}{\gamma\left(\mu^2 - \bar{L}^4\gamma^2\tau^2\alpha\right)}\log(\varepsilon_0/\varepsilon)$$

bits are sent, we ensure $\mathbf{E}\left(f(x_k) - f^\star\right) \leq \varepsilon$. *Here* B *is the number of bits required to encode a single vector entry and* $\mathbb{E}\{\|Q(v)\|_0\} \leq c$.

Proof. From Theorem 2, we get $V_k \leq \rho^k \varepsilon_0$ with $V_k = \mathbf{E}\left(f(x_k) - f^\star\right)$, or equivalently

$$\left(1 - \mu\gamma + \bar{L}^4\gamma^3\tau^2\frac{\alpha}{\mu}\right)^{\frac{k}{1+2\tau}} \varepsilon_0 \leq \varepsilon.$$

Since $-1/\log(1-x) \leq 1/x$ for $0 < x \leq 1$ and $\rho \in (0,1)$, we reach the upper bound of $k^\star$. In addition, assume that the number of non-zero elements is at most c. Therefore, the number of bits required to code the vector is at most $(\log_2 d + B)\, c$ bits in each iteration, where B is the number bits required to encode a single vector entry. Hence, we reach the upper bound of $B^\star$. □

I Proof of Theorem 3

Define $g_k = \sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i})$. Let us introduce three main lemmas which are instrumental in our main analysis.

Lemma 5. *The sequence* $\{x_k\}$ *generated by* (5) *satisfies*

$$\|\nabla f(x_k) - g_k\|^2 \leq \bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \|Q(g_j)\|^2,$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$ *and* $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$.

Proof. Following the proof in Lemma 4 yields the result. □

Lemma 6. *The sequence* $\{x_k\}$ *generated by* (5) *satisfies*

$$\mathbf{E}\,\|\nabla f(x_k) - Q(g_k)\|^2 \leq 2\left(1 + \beta(1+\theta)\right)\mathbf{E}\,\|\nabla f(x_k) - g_k\|^2 + 2\beta(1+1/\theta)\mathbf{E}\|\nabla f(x_k)\|^2,$$

where $\theta > 0$.

Proof. We start by deriving the upper bound of $\mathbf{E}\,\|g_k - Q(g_k)\|^2$. By the property stating that $\mathbf{E}\|Q(v) - v\|^2 \leq \beta\|v\|^2$ and by the fact that $\nabla f(x^\star) = 0$, we have:

$$\begin{aligned}\mathbf{E}\,\|g_k - Q(g_k)\|^2 &\leq \beta\|g_k\|^2 \\ &\leq \beta(1+\theta)\|g_k - \nabla f(x_k)\|^2 + \beta(1+1/\theta)\|\nabla f(x_k)\|^2,\end{aligned}$$

where the last inequality derives from the fact that $\|x+y\|^2 \leq (1+\theta)\|x\|^2 + (1+1/\theta)\|y\|^2$ for $x, y \in \mathbb{R}^d$ and $\theta > 0$.

Now, we are ready to derive the upper bound of $\mathbf{E}\,\|\nabla f(x_k) - Q(g_k)\|^2$. By the fact that $\|\sum_{i=1}^N x_i\|^2 \leq N\sum_{i=1}^N \|x_i\|^2$ for $x_i \in \mathbb{R}^d$ and $N \in \mathbb{N}$, we have

$$\|\nabla f(x_k) - Q(g_k)\|^2 \leq 2\,\|\nabla f(x_k) - g_k\|^2 + 2\,\|g_k - Q(g_k)\|^2.$$

Taking the expectation over the randomness and plugging the upper bound of $\mathbf{E}\left\|g_k - Q(g_k)\right\|^2$ into the result yield

$$\begin{aligned}\mathbf{E}\left\|\nabla f(x_k) - Q(g_k)\right\|^2 &\leq 2\mathbf{E}\left\|\nabla f(x_k) - g_k\right\|^2 + 2\mathbf{E}\left\|g_k - Q(g_k)\right\|^2 \\ &\leq 2\left(1+\beta(1+\theta)\right)\mathbf{E}\left\|\nabla f(x_k) - g_k\right\|^2 + 2\beta(1+1/\theta)\mathbf{E}\|\nabla f(x_k)\|^2.\end{aligned}$$

□

Lemma 7. *Suppose that non-negative sequences $\{V_k\}, \{w_k\}$, and $\{\Theta_k\}$ satisfying the following inequality*

$$V_{k+1} \leq V_k - a\Theta_k - bw_k + c\sum_{j=k-\tau}^{k} w_j, \tag{10}$$

where $a, b, c > 0$. Further suppose that $b - c(\tau+1) \geq 0$ and $w_k = 0$ for $k < 0$. Then,

$$\frac{1}{K+1}\sum_{k=0}^{K}\Theta_k \leq \frac{1}{a}\frac{1}{K+1}(V_0 - V_{K+1}).$$

Proof. Summing (10) from $k = 0$ to $k = K$ yields

$$\sum_{k=0}^{K} V_{k+1} \leq \sum_{k=0}^{K} V_k - a\sum_{k=0}^{K}\Theta_k - b\sum_{k=0}^{K} w_k + c\sum_{k=0}^{K}\sum_{j=k-\tau}^{k} w_j,$$

or equivalently due to the telescopic series

$$\begin{aligned}a\sum_{k=0}^{K}\Theta_k &\leq (V_0 - V_{K+1}) - b\sum_{k=0}^{K} w_k + c\sum_{k=0}^{K}\sum_{j=k-\tau}^{k} w_j \\ &= (V_0 - V_{K+1}) - b\sum_{k=0}^{K} w_k \\ &\quad + c(w_{-\tau} + w_{-\tau+1} + \ldots + w_0) \\ &\quad + c(w_{-\tau+1} + w_{-\tau+2} + \ldots + w_0 + w_1) + \ldots \\ &\quad + c(w_{-\tau+K} + w_{-\tau+K+1} + \ldots + w_0 + w_1 + \ldots + w_K) \\ &\leq (V_0 - V_{K+1}) - b\sum_{k=0}^{K} w_k + c(\tau+1)\sum_{k=0}^{K} w_k \\ &\leq V_0 - V_{K+1},\end{aligned}$$

where the second inequality comes from the fact that $w_k \geq 0$ for $k \geq 0$. In addition, the last inequality follows from the assumption that $b - c(\tau+1) \geq 0$. Then, we obtain the result. □

Now, we are ready to derive the convergence rate. From the definition of the Lipschitz continuity of the gradient of the function f, we have

$$\begin{aligned}f(x_{k+1}) - f^\star &\leq f(x_k) - f^\star - \gamma\left\langle\nabla f(x_k), Q(g_k)\right\rangle + \frac{\gamma^2\bar{L}}{2}\left\|Q(g_k)\right\|^2 \\ &\leq f(x_k) - f^\star - \frac{\gamma}{2}\left\|\nabla f(x_k)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\gamma^2\bar{L}}{2}\right)\left\|Q(g_k)\right\|^2 \\ &\quad + \frac{\gamma}{2}\left\|\nabla f(x_k) - Q(g_k)\right\|^2,\end{aligned}$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$ and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. The last inequality derives from the fact that $2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - \|x-y\|^2$ for any $x, y \in \mathbb{R}^d$. Next, taking the expectation over the randomness, and then plugging the inequality from Lemma 5 and 6 yield

$$\mathbf{E}f(x_{k+1}) - f^\star \leq \mathbf{E}f(x_k) - f^\star - \alpha_1\mathbf{E}\left\|\nabla f(x_k)\right\|^2 - \alpha_2\mathbf{E}\left\|Q(g_k)\right\|^2 + \alpha_3\sum_{j=k-\tau}^{k-1}\mathbf{E}\left\|Q(g_j)\right\|^2,$$

where

$$\alpha_1 = \gamma/2 - \gamma\beta(1+1/\theta)$$
$$\alpha_2 = \gamma/2 - \bar{L}\gamma^2/2$$
$$\alpha_3 = \gamma\left(1+\beta(1+\theta)\right)\bar{L}^2\gamma^2\tau,$$

and $\bar{L} = L\sqrt{m(1+\Delta)}$ and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. Next, we apply Lemma 7 with $V_k = \mathbf{E}f(x_k) - f^\star, \Theta_k = \mathbf{E}\|\nabla f(x_k)\|^2, w_k = \mathbf{E}\|Q(g_k)\|^2, a = \alpha_1, b = \alpha_2$, and $c = \alpha_3$. Notice that $\|Q(g_k)\| = \|x_{k+1} - x_k\|/\gamma$, which implies that $\|Q(g_k)\| = 0$ if $k < 0$. Therefore,

$$\frac{1}{K+1}\sum_{k=0}^{K}\Theta_k \leq \frac{1}{a}\frac{1}{K+1}(V_0 - V_{K+1}),$$

which means that

$$\min_{k\in[0,K]}\mathbf{E}\|\nabla f(x_k)\|^2 \leq \frac{1}{a}\frac{1}{K+1}\left(f(x_0) - f^\star\right) - \frac{1}{a}\frac{1}{K+1}\left(\mathbf{E}f(x_k) - f^\star\right).$$

To ensure the validity of the result, we must determine γ and β to satisfy three conditions, i.e. $a > 0$, $b > 0$ and $b - c(\tau+1) \geq 0$. The first criterion implies that $\beta < 1/\left(2(1+1/\theta)\right)$, and the last two criteria yield the admissible range of the step size γ. The second criterion implies that $\gamma < 1/\bar{L}$, and the equivalence of the last criterion is

$$\frac{1}{2} - \frac{\bar{L}\gamma}{2} - \left(1+\beta(1+\theta)\right)\bar{L}^2\tau(\tau+1)\gamma^2 \geq 0.$$

Therefore, let $\gamma = \frac{1}{\bar{L}(1+\omega)}$ where $\omega > 0$, and plugging the expression into the inequality yields

$$\omega^2 + \omega - 2\psi \geq 0,$$

where $\psi = \left(1+\beta(1+\theta)\right)\tau(\tau+1)$. Therefore, $\omega \geq \left(-1+\sqrt{1+8\psi}\right)/2$, and

$$\gamma < \frac{1}{\sqrt{1+8\left(1+\beta(1+\theta)\right)\tau(\tau+1)}}\frac{2}{\bar{L}}.$$

J Proof of Theorem 4

Since the component functions are convex and have L-Lipschitz continuous gradients,

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L\langle\nabla f_i(x) - \nabla f_i(y), x - y\rangle \qquad \forall x, y \in \mathbb{R}^d. \tag{11}$$

By Young's inequality,

$$\begin{aligned}\|\nabla f_i(x_k)\|^2 &\leq (1+\theta)\|\nabla f_i(x_k) - \nabla f_i(x^\star)\|^2 + (1+1/\theta)\|\nabla f_i^*\|^2 \\ &\leq (1+\theta)L\langle\nabla f_i(x_k) - \nabla f_i(x^\star), x_k - x^\star\rangle + (1+1/\theta)\|\nabla f_i(x^\star)\|^2\end{aligned} \tag{12}$$

We use the distance between the iterates $\{x_k\}_{k\in\mathbb{N}}$ and the optimum $x^\star$ to analyze the convergence:

$$\begin{aligned}\|x_{k+1} - x^\star\|^2 &= \|x_k - x^\star\|^2 - 2\gamma_k\left\langle\sum_{i=1}^{m}Q(\nabla f_i(x_k)), x_k - x^\star\right\rangle \\ &\quad + \gamma_k^2\left\|\sum_{i=1}^{m}Q(\nabla f_i(x_k))\right\|^2 \\ &\leq \|x_k - x^\star\|^2 - 2\gamma_k\left\langle\sum_{i=1}^{m}Q(\nabla f_i(x_k)), x_k - x^\star\right\rangle \\ &\quad + \gamma_k^2\sigma_k\sum_{i=1}^{m}\|Q(\nabla f_i(x_k))\|^2,\end{aligned}$$

where the second inequality comes from Lemma 2. Notice that $\mathbf{E}\left\|Q(\nabla f_i(x_k))\right\|^2 \leq \alpha \mathbf{E}\|\nabla f_i(x_k)\|^2$, since all machines have the same quantizers with the same parameters. Therefore, taking the expectation over all random variables yields

$$\begin{aligned}
\mathbf{E}\|x_{k+1}-x^\star\|^2 &= \mathbf{E}\|x_k-x^\star\|^2 - 2\gamma_k \mathbf{E}\left\langle \nabla f(x_k), x_k - x^\star\right\rangle \\
&\quad + \gamma_k^2 \sigma_k \sum_{i=1}^m \mathbf{E}\left\|Q(\nabla f_i(x_k))\right\|^2 \\
&\leq \mathbf{E}\|x_k-x^\star\|^2 - 2\gamma_k \mathbf{E}\left\langle \nabla f(x_k), x_k - x^\star\right\rangle \\
&\quad + \gamma_k^2 \sigma_k \sum_{i=1}^m \alpha \mathbf{E}\left\|\nabla f_i(x_k)\right\|^2 \\
&\leq \mathbf{E}\|x_k-x^\star\|^2 - 2\gamma_k \mathbf{E}\left\langle \nabla f(x_k), x_k - x^\star\right\rangle \\
&\quad + \gamma_k^2 \sigma_k \alpha \sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x_k)\right\|^2 \\
&\leq \mathbf{E}\|x_k-x^\star\|^2 - 2\gamma_k \mathbf{E}\left\langle \nabla f(x_k) - \nabla f(x^\star), x_k - x^\star\right\rangle \\
&\quad + \gamma_k^2 \sigma_k \alpha L(1+\theta) \mathbf{E}\langle \nabla f(x_k) - \nabla f(x^\star), x_k - x^\star\rangle \\
&\quad + \gamma_k^2 \sigma_k \alpha (1+1/\theta) \sum_{i=1}^m \mathbf{E}\|\nabla f_i(x^\star)\|^2,
\end{aligned}$$

where the last inequality comes from (12), $\nabla f(x) = \sum_{i=1}^m \nabla f_i(x)$, and $\nabla f(x^\star) = 0$. Now, let $\gamma_k = 1/\left(L\alpha(1+\theta)\sigma_k\right)$. Then, by strong convexity of f, we have:

$$\begin{aligned}
\mathbf{E}\|x_{k+1}-x^\star\|^2 &\leq \mathbf{E}\|x_k-x^\star\|^2 - \gamma_k \mathbf{E}\left\langle \nabla f(x_k) - \nabla f(x^\star), x_k - x^\star\right\rangle \\
&\quad + \gamma_k \frac{1}{\theta L} \sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2 \\
&\leq \rho_k \mathbf{E}\|x_k-x^\star\|^2 + \gamma_k \frac{1}{\theta L} \sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2,
\end{aligned}$$

where $\rho_k = 1-\mu\gamma_k$. Define $\rho_{\max} \in (0,1)$ and $\gamma_{\max}$ such that $\rho_k \leq \rho_{\max}$ and $\gamma_k \leq \gamma_{\max}\ \forall k$. Then

$$\|x_{k+1}-x^\star\|^2 \leq \rho_{\max}\|x_k-x^\star\|^2 + e_k$$

where

$$e_k = \gamma_{\max} \frac{1}{\theta L} \sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2.$$

Consequently, $V_k \leq \rho_{\max}^k V_0 + \bar{e}$ where $\bar{e} = e/(1-\rho_{\max})$.

If, instead, we use $\gamma_k = \gamma = 1/(L\alpha(1+\theta)\sigma)$, then a similar argument yields that

$$\|x_k-x^\star\|^2 \leq (1-\mu\gamma)^k \|x_0-x^\star\|^2 + \frac{1}{\mu\theta L} \sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2.$$

K Convergence Result of Distributed Quantized Gradient Method for Convex Optimization

Theorem 8. *Consider the optimization problem* (1) *under Assumption 1 and 3. Suppose that* $\gamma = 1/\left(L\alpha(1+\theta)\sigma\right)$ *where* $\theta > 0$ *and* $\sigma = \min\left(\sqrt{m(1+\Delta_{\text{ave}})}, 1+\Delta_{\max}\right)$. *Then, the iterates* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* (4) *satisfy*

$$\mathbf{E}(f(\bar{x}_T) - f(x^\star)) \leq \frac{1}{\gamma_{\min}} \frac{1}{T} \|x_0-x^\star\|^2 + \frac{1}{\theta L} \sum_{i=1}^m \mathbf{E}\|\nabla f_i(x^\star)\|^2,$$

where $\bar{x}_T = \frac{1}{T}\sum_{k=0}^{T-1} x_k$.

Proof. Following the proof in Theorem 4, we reach:

$$\begin{aligned}\mathbf{E}\|x_{k+1}-x^\star\|^2 \le \mathbf{E}\|x_k-x^\star\|^2 &- 2\gamma\mathbf{E}\left\langle\nabla f(x_k)-\nabla f(x^\star),x_k-x^\star\right\rangle\\ &+\gamma^2\sigma\alpha L(1+\theta)\mathbf{E}\langle\nabla f(x_k)-\nabla f(x^\star),x_k-x^\star\rangle\\ &+\gamma^2\sigma\alpha(1+1/\theta)\sum_{i=1}^m\mathbf{E}\|\nabla f_i(x^\star)\|^2.\end{aligned}$$

Now, let $\gamma = 1/\left(L\alpha(1+\theta)\sigma\right)$. Then, we have:

$$\begin{aligned}\mathbf{E}\|x_{k+1}-x^\star\|^2 &\le \mathbf{E}\|x_k-x^\star\|^2-\gamma\mathbf{E}\left\langle\nabla f(x_k)-\nabla f(x^\star),x_k-x^\star\right\rangle\\ &\quad+\gamma\frac{1}{\theta L}\sum_{i=1}^m\mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2\\ &\le \mathbf{E}\|x_k-x^\star\|^2-\gamma\mathbf{E}(f(x_k)-f(x^\star))+\gamma\frac{1}{\theta L}\sum_{i=1}^m\mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2,\end{aligned}$$

where the second inequality derives from the convexity of f, i.e. $\langle\nabla f(x_k),x_k-x^\star\rangle \ge f(x_k)-f(x^\star)$. Denote $\bar{x}_T = \frac{1}{T}\sum_{k=0}^{T-1}x_k$. Due to the convexity of the objective function f, $f(\bar{x}_T) \le \frac{1}{T}\sum_{k=0}^{T-1}f(x_k)$. By the manipulation, we have

$$\begin{aligned}\mathbf{E}(f(x_T)-f(x^\star)) &\le \frac{1}{T}\sum_{k=0}^{T-1}\mathbf{E}(f(x_k)-f(x^\star))\\ &\le \frac{1}{T}\sum_{k=0}^{T-1}\frac{1}{\gamma}\left(\mathbf{E}\|x_k-x^\star\|^2-\mathbf{E}\|x_{k+1}-x^\star\|^2\right)+\frac{1}{\theta L}\sum_{i=1}^m\mathbf{E}\|\nabla f_i(x^\star)\|^2\\ &\le \frac{1}{\gamma}\frac{1}{T}\|x_0-x^\star\|^2+\frac{1}{\theta L}\sum_{i=1}^m\mathbf{E}\|\nabla f_i(x^\star)\|^2,\end{aligned}$$

where we reach the last inequality by the telescopic series and by the non-negativity of the Euclidean norm. □

L Proof of Theorem 5

Denote $g_k = \sum_{i=1}^m\nabla f_i(x_{k-\tau_k^i})$. Before deriving the convergence rate, we introduce an essential lemma for our main analysis.

Lemma 8. *Consider the IAG update* (8) *with the URQ according to Definition 1. Then,*

$$\mathbf{E}\left\|e_k\right\|^2 \le 2m\sigma\alpha L^2\gamma^2\bar{L}^2\tau^2\max_{s\in[k-2\tau,k]}\left\|x_s-x^\star\right\|^2+2m\alpha\gamma^2\bar{L}^2\tau^2\sum_{i=1}^m\left\|\nabla f_i(x^\star)\right\|^2$$

where $\sigma = \min\left(\sqrt{m(1+\Delta_{\text{ave}})},1+\Delta_{\max}\right)$, $g_k = \sum_{i=1}^m\nabla f_i(x_{k-\tau_k^i})$, $\bar{L} = L\sqrt{m(1+\Delta)}$, *and* $\Delta = \min(\Delta_{\text{ave}},\Delta_{\max})$.

Proof. Denote $g_k = \sum_{i=1}^m\nabla f_i(x_{k-\tau_k^i})$ and $e_k = \nabla f(x_k)-g_k$. Following the proof of Lemma 1 with $x = x_k$ and $y = x_{k-\tau_k^i}$ yields

$$\left\|e_k\right\|^2 \le \bar{L}^2\max_{i\in[1,m]}\left\|x_k-x_{k-\tau_k^i}\right\|^2,$$

where $\bar{L} = L\sqrt{m(1+\Delta)}$, and $\Delta = \min(\Delta_{\text{ave}}, \Delta_{\max})$. Next, notice that

$$\begin{aligned}
\|e_k\|^2 &\leq \bar{L}^2 \max_{i\in[1,m]} \tau_k^i \sum_{j=k-\tau_k^i}^{k-1} \|x_{j+1} - x_j\|^2 \\
&\leq \bar{L}^2 \tau \sum_{j=k-\tau}^{k-1} \|x_{j+1} - x_j\|^2 \\
&\leq \gamma^2 \bar{L}^2 \tau \sum_{j=k-\tau}^{k-1} \left\| \sum_{i=1}^{m} Q(\nabla f_i(x_{j-\tau_j^i})) \right\|^2,
\end{aligned}$$

where the second inequality follows from the bounded delay assumption, and the last inequality from (8). On the other hand,

$$\begin{aligned}
\mathbf{E}\left\| \sum_{i=1}^{m} Q(\nabla f_i(x_{j-\tau_j^i})) \right\|^2 &\leq \sigma \sum_{i=1}^{m} \mathbf{E}\left\| Q(\nabla f_i(x_{j-\tau_j^i})) \right\|^2 \\
&\leq \sigma\alpha \sum_{i=1}^{m} \left\| \nabla f_i(x_{j-\tau_j^i}) \right\|^2 \\
&\leq 2\sigma\alpha \sum_{i=1}^{m} \left\| \nabla f_i(x_{j-\tau_j^i}) - \nabla f_i(x^\star) \right\|^2 + 2\sigma\alpha \sum_{i=1}^{m} \|\nabla f_i(x^\star)\|^2 \\
&\leq 2\sigma\alpha \sum_{i=1}^{m} L^2 \left\| x_{j-\tau_j^i} - x^\star \right\|^2 + 2\sigma\alpha \sum_{i=1}^{m} \|\nabla f_i(x^\star)\|^2 \\
&\leq 2m\sigma\alpha L^2 \max_{s\in[j-\tau,j]} \|x_s - x^\star\|^2 + 2m\alpha \sum_{i=1}^{m} \|\nabla f_i(x^\star)\|^2,
\end{aligned}$$

where we reach the first inequality by Lemma 2 due to Assumption 3; the second inequality by the second property of Definition 1; the third inequality by $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the forth inequality by the Lipschitz continuity assumption for gradient of each f_i; and the last inequality by the bounded delay assumption. Hence, plugging this result into the upper bound of e_k yields

$$\mathbf{E}\|e_k\|^2 \leq 2m\sigma\alpha L^2\gamma^2\bar{L}^2\tau^2 \max_{s\in[k-2\tau,k]} \|x_s - x^\star\|^2 + 2m\alpha\gamma^2\bar{L}^2\tau^2 \sum_{i=1}^{m} \|\nabla f_i(x^\star)\|^2.$$

□

We now prove Theorem 5. From (8), we have

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 = \|x_k - x^\star\|^2 &- 2\gamma \left\langle \sum_{i=1}^{m} Q\left(\nabla f_i(x_{k-\tau_k^i})\right), x_k - x^\star \right\rangle \\
&+ \gamma^2 \left\| \sum_{i=1}^{m} Q\left(\nabla f_i(x_{k-\tau_k^i})\right) \right\|^2.
\end{aligned}$$

Taking the expectation over all the random variables yields

$$\begin{aligned}
\mathbf{E}\|x_{k+1} - x^\star\|^2 = \mathbf{E}\|x_k - x^\star\|^2 &- 2\gamma\mathbf{E} \left\langle \sum_{i=1}^{m} \nabla f_i(x_{k-\tau_k^i}), x_k - x^\star \right\rangle \\
&+ \gamma^2\mathbf{E} \left\| \sum_{i=1}^{m} Q\left(\nabla f_i(x_{k-\tau_k^i})\right) \right\|^2.
\end{aligned}$$

Using the second property in Definition 1 and Lemma 2 due to Assumption 3, we get

$$\begin{aligned}
\mathbf{E}\left\|x_{k+1}-x^\star\right\|^2 &\le \mathbf{E}\left\|x_k-x^\star\right\|^2 - 2\gamma\mathbf{E}\left\langle \sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i}), x_k-x^\star\right\rangle \\
&\quad + \gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i})\right\|^2 \\
&= \mathbf{E}\left\|x_k-x^\star\right\|^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x_k), x_k-x^\star\right\rangle \\
&\quad + 2\gamma\mathbf{E}\left\langle g_k-\nabla f(x_k), x_k-x^\star\right\rangle \\
&\quad + \gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i})\right\|^2 \\
&\le \mathbf{E}\left\|x_k-x^\star\right\|^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x_k), x_k-x^\star\right\rangle \\
&\quad + \mathbf{E}\left\|g_k-\nabla f(x_k)\right\|^2 + \gamma^2\mathbf{E}\left\|x_k-x^\star\right\|^2 \\
&\quad + (1+\theta)\gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i}) - \nabla f_i(x^\star)\right\|^2 \\
&\quad + (1+1/\theta)\gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2
\end{aligned}$$

where $\sigma = \min\left(\sqrt{m(1+\Delta_{\text{ave}})}, 1+\Delta_{\max}\right)$ $g_k = \sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i})$. The second inequality follows from Cauchy-Schwarz's inequality and from the fact that $\|x+y\|^2 \le (1+\theta)\|x\|^2 + (1+1/\theta)\|y\|^2$ for $x, y \in \mathbb{R}^d$ and $\theta > 0$. Due to the Lipschitz continuity assumption of ∇f_i, we get

$$\begin{aligned}
\mathbf{E}\left\|x_{k+1}-x^\star\right\|^2 &\le (1+\gamma^2)\mathbf{E}\left\|x_k-x^\star\right\|^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x_k), x_k-x^\star\right\rangle \\
&\quad + \mathbf{E}\left\|g_k-\nabla f(x_k)\right\|^2 \\
&\quad + (1+\theta)\gamma^2 m\alpha\sigma L^2\mathbf{E}\left\|x_{k-\tau_k^i}-x^\star\right\|^2 \\
&\quad + (1+1/\theta)\gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2.
\end{aligned}$$

It follows from Lemma 8 that

$$\begin{aligned}
\mathbf{E}\left\|x_{k+1}-x^\star\right\|^2 &\le (1+\gamma^2)\mathbf{E}\left\|x_k-x^\star\right\|^2 - 2\gamma\mathbf{E}\left\langle \nabla f(x_k), x_k-x^\star\right\rangle \\
&\quad + 2m\sigma\alpha L^2\gamma^2\bar{L}^2\tau^2 \max_{s\in[k-2\tau,k]}\left\|x_s-x^\star\right\|^2 + 2m\alpha\gamma^2\bar{L}^2\tau^2\sum_{i=1}^m\left\|\nabla f_i(x^\star)\right\|^2 \\
&\quad + (1+\theta)\gamma^2 m\alpha\sigma L^2 \max_{s\in[k-\tau,k]}\mathbf{E}\left\|x_s-x^\star\right\|^2 \\
&\quad + (1+1/\theta)\gamma^2\sigma\alpha\sum_{i=1}^m \mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2.
\end{aligned}$$

Due to the property of the strong convexity assumption of f, it holds for $x, y \in \mathbb{R}^d$ that

$$\langle \nabla f(x) - \nabla f(y), x-y\rangle \ge \mu\|x-y\|^2,$$

Using this inequality with $x = x_k, y = x^\star$ and notice that $\nabla f(x^\star) = 0$ yields

$$V_{k+1} \le pV_k + q\max_{s\in[k-2\tau,k]} V_s + e,$$

where

$$
\begin{aligned}
V_k &= \mathbf{E}\|x_k - x^\star\|^2 \\
p &= 1 - 2\mu\gamma + \gamma^2 \\
q &= 2m\sigma\alpha L^2\gamma^2\bar{L}^2\tau^2 + (1+\theta)\gamma^2 m\alpha\sigma L^2 \\
e &= \left(2m\alpha\gamma^2\bar{L}^2\tau^2 + (1+1/\theta)\gamma^2\sigma\alpha\right)\sum_{i=1}^{m}\mathbf{E}\left\|\nabla f_i(x^\star)\right\|^2.
\end{aligned}
$$

From Lemma 1 of [21], $p+q<1$ implies that

$$
\gamma < \frac{2\mu}{1 + m\sigma\alpha L^2\left(2\bar{L}^2\tau^2 + (1+\theta)\right)}.
$$

Then, this implies that $V_k \leq (p+q)^{k/(1+2\tau)}V_0 + e/(1-p-q)$.

M Proof of Theorem 6

Denote $\tilde{g}_k = \sum_{i=1}^m Q\left(\nabla f_i(x_{k-\tau_k^i})\right)$ and $g_k = \sum_{i=1}^m \nabla f_i(x_{k-\tau_k^i})$. Before deriving the convergence rate, we introduce the lemmas which are instrumental in our main analysis.

Lemma 9. *The sequence* $\{x_k\}$ *generated by* (8) *satisfies*

$$
\|g_k - \nabla f(x_k)\|^2 \leq \bar{L}^2\gamma^2\tau\sum_{j=k-\tau}^{k-1}\|\tilde{g}_j\|^2.
$$

Proof. Following the proof in Lemma 4 yields the result. □

Lemma 10. *The sequence* $\{x_k\}$ *generated by* (8) *under Assumption 3 satisfies*

$$
\|g_k - \tilde{g}_k\|^2 \leq \sigma\sum_{i=1}^{m}\left\|\nabla f_i(x_{k-\tau_k^i}) - Q\left(\nabla f_i(x_{k-\tau_k^i})\right)\right\|^2,
$$

where $\sigma = \min\left(\sqrt{m(1+\Delta_{\text{ave}})}, 1+\Delta_{\max}\right)$.

Proof. The proof arguments follow those in Lemma 2 with replacing $Q(\nabla f_i(x_k))$ with $\nabla f_i(x_{k-\tau_k^i}) - Q\left(\nabla f_i(x_{k-\tau_k^i})\right)$. Also, note that $\operatorname{supp}\left(\nabla f_i(x_{k-\tau_k^i}) - Q\left(\nabla f_i(x_{k-\tau_k^i})\right)\right) \subset \operatorname{supp}(\nabla f_i(x_{k-\tau_k^i}))$, and Assumption 3 implies that $\operatorname{supp}(\nabla f_i(x_{k-\tau_k^i}))$ can be computed from the data directly. □

Lemma 11. *The sequence* $\{x_k\}$ *generated by* (8) *under Assumption 3 and 4 satisfies*

$$
\mathbf{E}\|\tilde{g}_k - \nabla f(x_k)\|^2 \leq 2\bar{L}^2\gamma^2\tau\sum_{j=k-\tau}^{k-1}\mathbf{E}\|\tilde{g}_j\|^2 + 2\beta\sigma mC^2.
$$

Proof. By the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, we have:

$$
\begin{aligned}
\|\tilde{g}_k - \nabla f(x_k)\|^2 &\leq 2\|g_k - \nabla f(x_k)\|^2 + 2\|g_k - \tilde{g}_k\|^2 \\
&\leq 2\bar{L}^2\gamma^2\tau\sum_{j=k-\tau}^{k-1}\|\tilde{g}_j\|^2 + 2\|g_k - \tilde{g}_k\|^2,
\end{aligned}
$$

where the last inequality follows from Lemma 9. Next, taking the expectation of the inequality from Lemma 10 over the randomness yields

$$
\mathbf{E}\|g_k - \tilde{g}_k\|^2 \leq \sigma\sum_{i=1}^{m}\mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i}) - Q\left(\nabla f_i(x_{k-\tau_k^i})\right)\right\|^2.
$$

Next, taking the expectation over the randomness yields

$$\begin{aligned}
\mathbf{E}\|\tilde{g}_k - \nabla f(x_k)\|^2 &\leq 2\bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \mathbf{E}\|\tilde{g}_j\|^2 + 2\mathbf{E}\|g_k - \tilde{g}_k\|^2 \\
&\leq 2\bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \mathbf{E}\|\tilde{g}_j\|^2 + 2\sigma \sum_{i=1}^{m} \mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i}) - Q\left(\nabla f_i(x_{k-\tau_k^i})\right)\right\|^2 \\
&\leq 2\bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \mathbf{E}\|\tilde{g}_j\|^2 + 2\beta\sigma \sum_{i=1}^{m} \mathbf{E}\left\|\nabla f_i(x_{k-\tau_k^i})\right\|^2 \\
&\leq 2\bar{L}^2\gamma^2\tau \sum_{j=k-\tau}^{k-1} \mathbf{E}\|\tilde{g}_j\|^2 + 2\beta\sigma m C^2,
\end{aligned}$$

where we reach the third inequality by the property of the URQ stating that $\mathbf{E}\|Q(v)-v\|^2 \leq \beta\mathbf{E}\|v\|^2$, and the last inequality by Assumption 4. □

Lemma 12. *Suppose that non-negative sequences $\{V_k\}, \{w_k\}$, and $\{\Theta_k\}$ satisfying the following inequality*

$$V_{k+1} \leq V_k - a\Theta_k - bw_k + c\sum_{j=k-\tau}^{k} w_j + e, \tag{13}$$

where $a, b, c, e > 0$. Further suppose that $b - c(\tau+1) \geq 0$ and $w_k = 0$ for $k < 0$. Then,

$$\frac{1}{K+1}\sum_{k=0}^{K}\Theta_k \leq \frac{1}{a}\frac{1}{K+1}(V_0 - V_{K+1}) + \frac{1}{a}e.$$

Proof. Following the proof in Lemma 7 yields the result. □

Now, we are ready to derive the convergence rate. From the Lipschitz continuity assumption of the gradient of f,

$$\begin{aligned}
f(x_{k+1}) - f^\star &\leq f(x_k) - f^\star - \gamma\left\langle \nabla f(x_k), \tilde{g}_k\right\rangle + \frac{\bar{L}\gamma^2}{2}\left\|\tilde{g}_k\right\|^2 \\
&= f(x_k) - f^\star - \frac{\gamma}{2}\left\|\nabla f(x_k)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\bar{L}\gamma^2}{2}\right)\left\|\tilde{g}_k\right\|^2 + \frac{\gamma}{2}\left\|\tilde{g}_k - \nabla f(x_k)\right\|^2,
\end{aligned}$$

where the equality derives from the fact that $2\langle x, y\rangle = \|x\|^2 + \|y\|^2 - \|x-y\|^2$ for $x, y \in \mathbb{R}^d$. Taking the expectation over the randomness and using Lemma 11 yields

$$\begin{aligned}
\mathbf{E}f(x_{k+1}) - f^\star &\leq \mathbf{E}f(x_k) - f^\star - \frac{\gamma}{2}\mathbf{E}\left\|\nabla f(x_k)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\bar{L}\gamma^2}{2}\right)\mathbf{E}\left\|\tilde{g}_k\right\|^2 + \frac{\gamma}{2}\mathbf{E}\left\|\tilde{g}_k - \nabla f(x_k)\right\|^2 \\
&\leq \mathbf{E}f(x_k) - f^\star - \frac{\gamma}{2}\mathbf{E}\left\|\nabla f(x_k)\right\|^2 - \left(\frac{\gamma}{2} - \frac{\bar{L}\gamma^2}{2}\right)\mathbf{E}\left\|\tilde{g}_k\right\|^2 \\
&\quad + \bar{L}^2\gamma^3\tau \sum_{j=k-\tau}^{k-1} \mathbf{E}\|\tilde{g}_j\|^2 + \gamma\beta\sigma m C^2
\end{aligned}$$

Next, applying Lemma (12) with $V_k = \mathbf{E}f(x_k) - f^\star, \Theta_k = \mathbf{E}\|\nabla f(x_k)\|^2, w_k = \mathbf{E}\|\tilde{g}_k\|^2, e = \gamma\beta\sigma m C^2, a = \gamma/2, b = \gamma/2 - \bar{L}\gamma^2/2$, and $c = \bar{L}^2\tau\gamma^3$ yields the result.

$$\min_{k\in[0,K]} \mathbf{E}\|\nabla f(x_k)\|^2 \leq \frac{1}{a}\frac{1}{K+1}\left(f(x_0) - f^\star\right) - \frac{1}{a}\frac{1}{K+1}\left(f(x_k) - f^\star\right) + \frac{1}{a}e.$$

Note that $w_k = 0$ for $k < 0$ since $\mathbf{E}\|\tilde{g}_k\|^2 = \mathbf{E}\|x_{k+1} - x_k\|^2/\gamma^2$. Lastly, we need to find the admissible range of the step-size which guarantees the convergence. The following criteria must be

satisfied: $b > 0$ and $b - c(\tau + 1) \geq 0$. The first criterion implies that $\gamma < 1/\bar{L}$. The second criterion implies that

$$\frac{\gamma}{2} - \frac{\bar{L}\gamma^2}{2} - \bar{L}^2\tau(\tau + 1)\gamma^3 \geq 0.$$

Lastly, let $\gamma = 1/(\bar{L} + \omega)$ for $\omega > 0$ and plugging the expression into the result yields

$$\omega^2 + \bar{L}\omega - 2\bar{L}^2\tau(\tau + 1) \geq 0,$$

and therefore

$$\omega \geq \left(-1 + \sqrt{1 + 8\tau(\tau + 1)}\right) \frac{\bar{L}}{2}.$$

Thus, we can conclude that the admissible range of the step-size is

$$\gamma < \frac{1}{1 + \sqrt{1 + 8\tau(\tau + 1)}} \frac{2}{\bar{L}}.$$