2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

Gradient compression for communication-limited convex optimization

Sarit Khirirat, Mikael Johansson, Dan Alistarh

***Abstract*— Data-rich applications in machine-learning and control have motivated an intense research on large-scale optimization. Novel algorithms have been proposed and shown to have optimal convergence rates in terms of iteration counts. However, their practical performance is severely degraded by the cost of exchanging high-dimensional gradient vectors between computing nodes. Several gradient compression heuristics have recently been proposed to reduce communications, but few theoretical results exist that quantify how they impact algorithm convergence. This paper establishes and strengthens the convergence guarantees for gradient descent under a family of gradient compression techniques. For convex optimization problems, we derive admissible step sizes and quantify both the number of iterations and the number of bits that need to be exchanged to reach a target accuracy. Finally, we validate the performance of different gradient compression techniques in simulations. The numerical results highlight the properties of different gradient compression algorithms and confirm that fast convergence with limited information exchange is possible.**

I. INTRODUCTION

The traditional complexity measure for optimization, *iteration complexity*, is often a good indicator of the total running time of the algorithm. The iteration complexity of first-order methods for convex optimization is by now well-established: fundamental lower bounds have been derived [5], [8] and optimization algorithms with order-optimal convergence rates have been developed [7], [8]. When the decision vectors become large, however, the communication cost for exchanging gradients between computing nodes escalates. Simply reducing the number of iterations is no longer sufficient. Rather, it becomes essential to try to minimize the total amount of communication required to reach an ϵ-optimal solution. Although the *communication complexity* of convex optimization has received some attention in the past [1], [2], [3], [4], the concept is much less well understood than iteration complexity. Lower bounds on communication complexity exist only for special classes of functions and under restrictive assumptions; and optimal algorithms are yet unknown, even for unconstrained convex optimization.

A natural way to limit the amount of data exchanged is to compress the information exchanged between computing nodes. To this end, several lossy gradient compression heuristics have recently been proposed and empirically shown to be able to reduce the amount of communication. However, only a few theoretical results exist that quantify how they impact algorithm convergence (*e.g.*, [6], [9], [12]).

[1]S. Khirirat and M. Johansson are with the Department of Automatic Control, School of Electrical Engineering and ACCESS Linnaeus Center, Royal Institute of Technology (KTH), Stockholm, Sweden. Emails: {*sarit@kth.se, mikaelj@kth.se*}. D. Alistarh is with Institute of Science and Technology (IST) Austria, Vienna, Austria. Email: {*dan.alistarh@ist.ac.at*}.

In this paper, we study several gradient compression techniques and quantify their impact on the iteration and communication complexity of the gradient descent algorithm for convex optimization. For each class of gradient compression algorithms, we derive valid inequalities which are useful for analyzing their effect on optimization algorithms. We then use these inequalities to establish iteration and communication complexity bound of the gradient descent algorithm operating on compressed gradients. In some cases, our results are the first in the literature; in other cases, our results extend and strengthen the current state-of-the art. We validate our theoretical findings in simulations, and demonstrate that the choice of the gradient compression technique can have a significant impact on the algorithm performance.

A. Notation

We let $\mathbb{N}$ and $\mathbb{N}_0$ be the set of natural numbers and the set of natural numbers including zero, respectively. The sign of $y \in \mathbb{R}$, $\mathrm{sgn}(y)$, is defined as $\mathrm{sgn}(y) = 1$ if $y > 0$, -1 if $y < 0$, and 0 otherwise. We let $\lceil x \rceil$ denote the least integer which is greater than or equal to x, $[0, T]$ be the set $\{0, 1, \ldots, T\}$, and $|A|$ be the cardinality of the set A. For $x \in \mathbb{R}^n$, $\|x\|$ and $\|x\|_1$ are the ℓ_2 norm and the ℓ_1 norm of x, respectively; x_i is the i^{th} coordinate of x; $\mathrm{supp}(x) = \{i : x_i \neq 0\}$; and the sign vector of $x \in \mathbb{R}^n$ is defined as $\mathrm{sgn}(x)_i = \mathrm{sgn}(x_i)$. Finally, $\mathcal{U}(0,1)$ is the uniform distribution on the interval $[0,1]$ and $\mathcal{N}(0,1)$ is the zero-mean Gaussian distribution with unit variance.

II. PROBLEM FORMULATION

We consider convex optimization problems on the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x). \tag{1}$$

To facilitate the analysis, we make the following standard assumptions on the class of loss functions.

Assumption 1: The loss function $f : \mathbb{R}^n \to \mathbb{R}$ has L-Lipschitz continuous gradient, *i.e.* there exists $L > 0$ such that for all $x, y \in \mathbb{R}^n$

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

Assumption 2: The loss $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex, *i.e.* there exists $\mu > 0$ such that for all $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2. \tag{3}$$

The classical *gradient descent* (GD) for solving (1) is to form the sequence $\{x_k\}_{k \in \mathbb{N}}$ such that

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

given the initial point x_0 and fixed positive step size γ. Theoretical results (e.g., in [7], [8]) guarantee that GD reaches ϵ accuracy after at most $\mathcal{O}(1/\epsilon)$ iterations when f is convex and satisfies Assumption 1. If each entry of the gradient is coded using C bits, then representing the total vector requires nC bits, and an ϵ-optimal solution is obtained after $\mathcal{O}(nC/\epsilon)$ bits have been sent from the oracle (responsible for computing the gradient of the loss function) to the master (responsible for updating the decision vector).

To reduce the amount of communication from the oracle to the master, we consider various techniques for compressing the gradient vector and study the convergence of the iteration

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

As shown in Section IV, different quantizers Q result in messages of different size (in number of bits), but they also impact the number of iterations required to reach target accuracy. Our aim in this paper is to quantify the composite effect which the quantizer selection has on the total number of bits exchanged until an ϵ-optimal solution has been found.

III. GRADIENT COMPRESSION

We begin by introducing a generic compression operator which captures the key features of the most popular gradient compression schemes in the literature.

Definition 1: The operator $Q : \mathbb{R}^d \to \mathbb{R}^d$ is called a *bounded relative error quantizer (BREQ)* if for all $v \in \mathbb{R}^d$ and for some positive constants α, β

(a) $\langle Q(v), v\rangle \geq \alpha\|v\|^2$, and
(b) $\|Q(v)\|^2 \leq \beta\|v\|^2$.

Note that the first inequality is satisfied by any quantizer for which $\mathrm{sgn}(Q(v)) = \mathrm{sgn}(v)$. Moreover, conditions (a) and (b) imply that $\|Q(v) - v\|^2 \leq (1 - 2\alpha + \beta)\|v\|^2$, so the relative quantization error induced by Q is indeed bounded. When $\alpha = \beta = 1$, then $Q(v) = v$. Next, we give several examples of gradient compression schemes which are BREQs. The proof of all claims can be found in appendix.

A. Sparsification

One natural technique for gradient compression is *sparsification*, *i.e.* setting small vector elements to zero. This sparsification can either be done based on threshold on the magnitude of the vector elements, or by keeping a fixed number K of components. In this paper, we focus on the latter case according to the following definition.

Definition 2: The K-greedy quantizer $Q_G^K : \mathbb{R}^n \mapsto \mathbb{R}^n$ is

$$[Q_G^K(g)]_i = \begin{cases} [g]_{\pi(i)} & \text{if } i \leq K \\ 0 & \text{otherwise} \end{cases}$$

where π is a permutation of $\{1, \ldots, n\}$ such that $|g_{\pi(k)}| \geq |g_{\pi(k+1)}|$ for all $k \in \{1, \ldots, n-1\}$.

The case $K = 1$ has been treated by Nutini *et al.* [10]. A naive encoding of a vector processed by the K-greedy quantizer requires $K(\log_2(n) + b)$ bits: $\log_2(n)$ bits to represent each index and b bits to represent the corresponding entry of the K non-zero values.

Lemma 1: The $K-$greedy quantizer Q_G^K is a BREQ with $\alpha = K/n$ and $\beta = 1$. In addition, for any $g \in \mathbb{R}^n$,

$$(K/n)\|g\|^2 \leq \|Q_G^K(g)\|^2 \tag{5}$$

B. Quantization

Another approach to reduce the size of the gradient vector is to quantize the individual elements. At the extreme, one can consider three-level (ternary) quantization, where each vector element is quantized to the levels $\{-1, 0, 1\}$. The convergence of gradient descent with the ternary quantizer has been studied in [9]. One drawback with this quantizer is that the absence of magnitude information about the original gradient leads to a residual error in the gradient descent. To avoid this problem, one can rather code each element of g to $\{-\|g\|, 0, \|g\|\}$. We thus consider the following quantizer:

Definition 3: The ternary quantizer $Q_T : \mathbb{R}^n \mapsto \mathbb{R}^n$ is

$$[Q_T(g)]_i = \|g\|\mathrm{sgn}(g_i).$$

The required number of bits to encode the gradient by the ternary quantizer is $2n + b$: b bits to encode the norm of the vector, and 2 bits for each element to encode its sign.

Lemma 2: The ternary quantizer Q_T is BREQ with $\alpha = 1, \beta = |\mathrm{supp}(Q_T(g))| \leq n$.

C. Combined quantization and sparsification

Finally, one can also combine sparsification and quantization; the compressed gradient is then represented by its (uncompressed) magnitude and the sign of a few entries. Such quantizers have been recently proposed in, *e.g.*, [12], [14]. In this paper, we consider the following deterministic variant of the dynamic gradient quantizer introduced in [12].

Definition 4: The *dynamic gradient quantizer* $Q_D : \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined as

$$[Q_D(g)]_i = \begin{cases} \|g\|\mathrm{sgn}(g_i) & \text{if } i \in I(g) \\ 0 & \text{otherwise} \end{cases}$$

where $I(g)$ is the smallest subset of $\{1, \ldots, n\}$ such that

$$\sum_{i \in I(g)} |g_i| \geq \|g\|.$$

The dynamic gradient quantizer was analyzed in Alistarh *et al.* [12]. The dynamic quantizer requires $|I(g)|(\log_2(n) + 1) + b$ bits to encode the gradient.

Lemma 3: The dynamic quantizer Q_D is BREQ with $\alpha = 1, \beta = |I(g)| \leq \sqrt{n}$.

Lemma 3 is a slight extension of [12, Lemma F.1].

IV. CONVERGENCE ANALYSIS

This section establishes a unified convergence analysis of the gradient descent algorithm under BREQ compression in terms of both iteration counts and number of communicated bits. We thus consider

$$x_{k+1} = x_k - \gamma_k Q\left(\nabla f(x_k)\right), \tag{6}$$

where γ_k is a positive step size and Q is a generic BREQ introduced in Section III. Our first result is the following.

Theorem 1: Consider the problem (1) under Assumption 1 and 2. Suppose that $\gamma_k = \alpha/(\beta L)$. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) satisfy

$$f(x_k) - f^\star \leq \rho^k \left(f(x_0) - f^\star\right),$$

where $\rho = 1 - \alpha^2\mu/(2\beta L) \in (0,1)$.

Proof: See the appendix. ∎

Theorem 1 establishes linear convergence of gradient descent under BREQ compression, but implies a convergence rate penalty of $\alpha^2/2\beta$ compared to a similar analysis of the full gradient descent method. This penalty can be significant, and ranges from $1/2\sqrt{n}$ for the dynamic quantizer to $K^2/2n^2$ for the K-greedy quantizer. One limitation of our analysis is that it is based on upper bounds on the cardinality of the compressed vectors, although the actual cardinality of each compressed vector is known by the algorithm at runtime. The next result establishes convergence when we allow the step-size in the gradient descent to depend on the actual support set of the compressed vectors.

Theorem 2: Consider the problem (1) under Assumption 1 and 2. Suppose that $\gamma_k = \alpha_k/(\beta_k L)$. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) satisfy

$$f(x_k) - f^\star \leq (f(x_0) - f^\star) \prod_{t=0}^{k} \rho_t,$$

where $\rho_t = 1 - \alpha_t^2\mu/(2\beta_t L) \in (0,1)$.

Proof: See the appendix. ∎

Clearly, the convergence rate remains linear. We will demonstrate in Section V that time-varying step-sizes can lead to significant improvements in both iteration and communication complexity. The results extend to convex and non-convex optimization problems, as shown below.

Theorem 3 (Convex optimization): Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex and satisfy Assumption 1. Let $\{x_k\}$ be generated by (6) with $\gamma_k = \alpha_k/(\beta_k L)$. If there exits a positive constant R such that $\|x_k - x^\star\| \leq R$ for all k, then

$$f(x_k) - f^\star \leq \frac{1}{k}\frac{1}{\Omega}R^2,$$

where $\Omega = \min_{t\in[0,k-1]} \alpha_t^2/(2\beta_t L)$.

Proof: See the appendix. ∎

Theorem 4 (Non-convex optimization): Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy Assumption 1. If $\gamma_k = \alpha_k/(\beta_k L)$, then the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) satisfy

$$\min_{t\in[0,k-1]} \|\nabla f(x_t)\|^2 \leq \frac{1}{k}\frac{1}{\Omega_{\min}} \left(f(x_0) - f(x_k)\right),$$

where $\Omega_{\min} = \min_{t\in[0,k-1]} \alpha_t^2/(2\beta_t L)$.

Proof: See the appendix. ∎

Theorem 3 and 4 both imply that gradient descent with BREQ compression retains the sub-linear rate $\mathcal{O}(1/T)$, and like Theorem 2 the convergence factors are penalized by a term $\alpha^2/2\beta$ which depends on the accuracy of the compression algorithm. The results also allow us to estimate the iteration and communication complexity, as shown next.

Corollary 1: Consider the problem (1) under Assumption 1. Let c be the required number of bits to encode one compressed vector and denote $\Omega_k = \alpha_k^2/(2\beta_k)$. Suppose that $\gamma_k = \alpha_k/(\beta_k L)$. Given $\varepsilon_0 = f(x_0) - f(x_T)$, by running (6) for at most

$$T^\star = \frac{1}{\min_{k\in[0,T-1]}\Omega_k} L\varepsilon_0/\varepsilon,$$

iterations, under which $B^\star = \lceil cT^\star \rceil$ bits are sent, we ensure $\min_{k\in[0,T-1]} \|\nabla f(x_k)\|^2 \leq \varepsilon$.

Proof: Due to limited space, we omit the proof. ∎

A. Tighter Convergence of $K-$greedy quantizer

The greedy quantizer has a small value of α, which translates into a high convergence penalty in the unified analysis. However, by exploiting the lower bound on $\|Q(v)\|$ stated in Lemma 6, we can give convergence guarantees that are of the same order of magnitude as the other BREQs.

Theorem 5: Let $f : \mathbb{R}^n \to \mathbb{R}$, and consider the iterates $\{x_k\}$ generated by (6) under K-greedy gradient compression. If f satisifies Assumption 1 and $\gamma = 1/L$, then

$$\min_{t\in[0,k-1]} \|\nabla f(x_t)\|^2 \leq \frac{1}{k}\frac{1}{\Omega} \left(f(x_0) - f^\star\right),$$

where $\Omega = K/(2nL)$. In addition, if f is convex, then

$$f(x_k) - f^\star \leq \frac{1}{k}\frac{1}{\Omega}R^2,$$

where $R^2 \geq \|x_k - x^\star\|$ for all k. If f is convex and satisfies both Assumption 1 and 2, and $\gamma = 1/L$, then

$$f(x_k) - f^\star \leq \rho^k (f(x_0) - f^\star),$$

where $\rho = 1 - KL/(2n\mu)$.

Analogously to the generic BREQ analysis, we can derive the following bound on the communication complexity of gradient descent under K-greedy sparsification of gradients.

Corollary 2: Consider the problem (1) under Assumption 1. Let c be the required number of bits to encode one compressed vector and denote $\Omega = K/(2n)$. Suppose that $\gamma_k = 1/L$. Given $\varepsilon_0 = f(x_0) - f(x_T)$, by running (6) with the $K-$greedy quantizer for at most

$$T^\star = \frac{1}{\Omega} L\varepsilon_0/\varepsilon,$$

iterations, under which $B^\star = \lceil cT^\star \rceil$ bits are sent, we ensure $\min_{k\in[0,T-1]} \|\nabla f(x_k)\|^2 \leq \varepsilon$.

Proof: Due to limited space, we omit the proof. ∎

The iteration and communication complexities in Corollary 1 and 2 are summarized in Table I.

V. EXPERIMENTAL RESULTS

We evaluated the performance of GD with different gradient compression techniques on a least-squares problem in Julia. This problem is on the form (1) with $f(x) = \frac{1}{2}\|Ax - b\|^2$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Test instances with $m = 1000$ and $n = 800$ were created as follows: each element of A was drawn from $\mathcal{U}(0,1)$ and each element of b was set to be the sign of a random number drawn from $\mathcal{N}(0,1)$. We normalized each row of A by its Euclidean norm and computed the Lipschitz constant as $L = \lambda_{\max}(A^T A)$. The

Quantizer	$\gamma_k^\star$	$T^\star$	$B^\star$
$K-$greedy quantizer	$\frac{1}{L}$	$\frac{n}{K}\nu$	$\lceil n(\log_2(n)+b)\nu \rceil$
Ternary quantizer	$\frac{1}{K_k L}$	$K_{\max}\nu$	$\lceil K_{\max}(2n+b)\nu \rceil$
Dynamic gradient quantizer	$\frac{1}{K_k L}$	$K_{\max}\nu$	$\lceil K_{\max} C_D \nu \rceil$

TABLE I

ITERATION AND COMMUNICATIONS COMPLEXITY OF COMPRESSED GD WHERE $C_D = K_{\max}(\log_2(n)+1)+b$, $K_k = |\mathrm{supp}(Q(\nabla f(x_k)))|$, $K_{\max} = \max_k K_k$, AND $\nu = 2L\varepsilon_0/\varepsilon$.

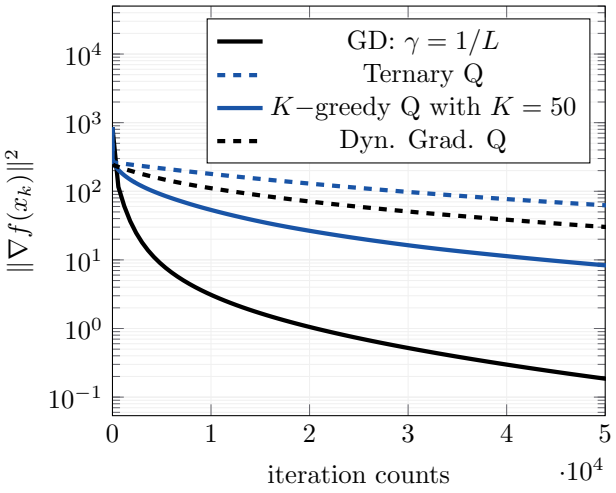
compressed gradient iterations were initialized from $x_0 = \mathbf{0}$, and we assumed that real numbers were represented by $b = 64$ bits. The step sizes under different gradient compression algorithms are tuned according to Table I.

From Figure 1(a), GD has the fastest convergence towards the optimum in terms of iteration counts. This is expected, since all gradient compression techniques introduce an information loss. However, GD with gradient compression tends to have better performance than GD in terms of the number of communicated bits; see Figure 1(b). The exception is the ternary quantizer, which is uniformly worse than its alternatives, possibly due to its small theoretically justified step size. Figure 1(b) indicates that GD with the dynamic gradient quantizer attains the best communication complexity among GD with other compression techniques.

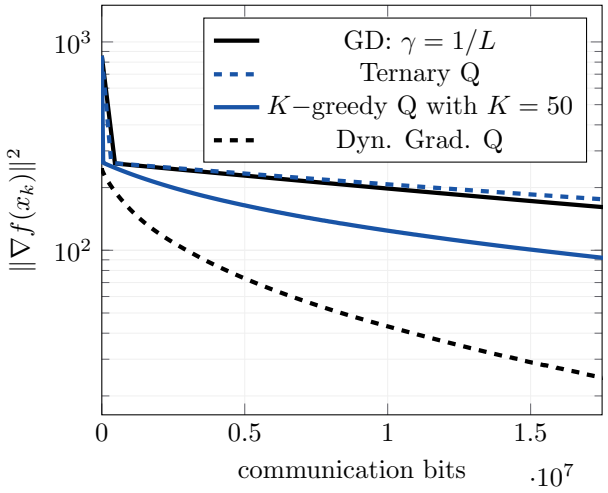
VI. CONCLUSIONS

Driven by the need to reduce communication in distributed optimization applications, we have investigated how various gradient compression schemes impact iteration and communication complexity of the gradient descent algorithm. We have presented a uniform analysis of gradient descent under several important gradient compression schemes from the literature. The results cover both convex, strongly convex, and non-convex loss functions and reveal how the quantization error impacts admissible step-sizes and affects the guaranteed convergence times. Upper bounds on the number of communicated bits to ensure a target level of suboptimality are also derived. Three important gradient compression schemes were studied in detail. For some combinations of compression algorithms and loss functions, our results are the first in the literature; for other combinations, our results extend and strengthen existing results in the literature. Our numerical results demonstrate how the dynamic gradient quantizer is able to outperform its alternatives in terms of both iteration and communication complexity.

Future work includes analyzing bi-directional compression (*i.e.* both of the iterate vector and the gradients), consideration of randomized quantizers, and the use of gradient compression in other classes of optimization algorithms.



(a)



(b)

Fig. 1. The performance of compressed GD and GD w.r.t. iteration counts (Figure 1(a)) and the number of communicated bits (Figure 1(b)).

VII. ACKNOWLEDGMENTS

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

APPENDIX I
DERIVATION OF BREQ PARAMETERS

$K-$greedy quantizer

Clearly,

$$\|Q_G^K(g)\|^2 = \sum_{i\in I(g)} |g_i|^2 \leq \|g\|_2^2$$

so $\beta = 1$ is a valid estimate. To estimate α, we write $Q_G^K(g) = \sum_{i\in I(g)} e_i g_i$, where g_i is the i^{th} element of $g \in \mathbb{R}^n$; $I(g)$ collects K indices corresponding elements of g with largest absolute value; and $e_i \in \{0,1\}^n$ with only 1 at component $i \in I(g)$ and zeroes elsewhere. Then

$$\langle g, Q_G^K(g)\rangle = \langle g, \sum_{i\in I(g)} e_i g_i\rangle = \sum_{i\in I(g)} |g_i|^2 = \|Q_G^K(g)\|^2.$$

Introducing $I^c(g)$ as the complement of the set $I(g)$, we note that for every $j \in I^c(g)$,

$$|g_j|^2 \leq \min_{i\in I(g)} |g_i|^2 \leq \frac{1}{K}\sum_{i\in I(g)} |g_i|^2.$$

As $|I(g)| = K$ and $|I^c(g)| = n - K$

$$\begin{aligned}\|g\|^2 &= \sum_{i\in I(g)} |g_i|^2 + \sum_{j\in I^c(g)} |g_j|^2 \\ &\leq (1 + \tfrac{n-K}{K}) \sum_{i\in I(g)} |g_i|^2 \\ &= \tfrac{n}{K}\|Q_G^K(g)\|^2.\end{aligned}$$

which implies the inequality (5) and that

$$\langle g, Q_G^K(g)\rangle \geq (K/n)\|g\|^2$$

Hence, $\alpha = K/n$ is a valid estimate. ■

Ternary quantizer

Since $\|g\|_1 = \sum_{i=1}^n g_i \mathrm{sgn}(g_i)$ and $\|g\|_1 \geq \|g\|$

$$\langle g, Q_T(g)\rangle = \|g\|\|g\|_1 \geq \|g\|^2$$

so $\alpha = 1$ is valid. Next, $\|Q_T(g)\|^2 = |\mathrm{supp}(Q_T(g))|\cdot\|g\|^2 \leq n\|g\|^2$, confirms the bounds for β. ■

Dynamic gradient quantizer

By Definition 4, we have

$$\begin{aligned}\langle g, Q_D(g)\rangle &= \langle g, \|g\| \sum_{i\in I(g)} \mathrm{sgn}(g_i) e_i\rangle \\ &= \|g\| \sum_{i\in I(g)} g_i \mathrm{sgn}(g_i) = \|g\| \sum_{i\in I(g)} |g_i|\end{aligned}$$

By the construction of $I(g)$, we thus conclude that

$$\langle g, Q_D(g)\rangle \geq \|g\|^2.$$

In addition, $\|Q_D(g)\|^2 = |I(g)| \cdot \|g\|^2$. From [12, Lemma F.1], we know that $|I(g)| \leq \sqrt{n}$. This confirms the proposed bounds on α and β. ■

APPENDIX II
PROOFS OF MAIN RESULTS

Lemma 4

This lemma establishes our unified theoretical results of the compressed gradient descent with the fixed step-size.

Lemma 4: Consider the optimization problem (1) under Assumption 1. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) with the BREQs according to Definition 1 satisfy

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

Proof: From the definition of the Lipschitz continuity of ∇f and (6), we have:

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

Applying two inequalities of BREQ from Definition 1 into the main result completes the proof. ■

Proof of Theorem 1

By the strong convexity assumption of f,

$$\|\nabla f(x_k)\|^2 \geq 2\mu\left(f(x_k) - f^\star\right).$$

Assume that $\gamma_k = \gamma$ such that $\gamma = \alpha/(\beta L)$. Applying this inequality into the one in Lemma 4 yields

$$f(x_{k+1}) - f^\star \leq \rho\left(f(x_k) - f^\star\right),$$

where $\rho_k = 1 - \Gamma/\kappa$, $\Gamma = \alpha^2/(2\beta)$ and $\kappa = L/\mu$. Suppose that $\rho \in (0, 1)$. Then, by the recursion of the inequality, we obtain the result. ■

Lemma 5

This lemma extends our earlier results to compressed gradient descent with time-varying step-sizes.

Lemma 5: Consider the optimization problem (1) under Assumption 1. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) with the BREQs according to Definition 1 satisfy

$$f(x_{k+1}) \leq f(x_k) - \gamma_k \left(\alpha_k - \frac{L\beta_k\gamma_k}{2}\right)\|\nabla f(x_k)\|^2.$$

Proof: From the definition of the Lipschitz continuity of ∇f and (6), we have:

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

Note that BREQ parameters are time-varying, i.e. we have α_k, β_k. Applying two inequalities of BREQ from Definition 1 into the main result completes the proof. ■

Proof of Theorem 2

By the strong convexity of f,

$$\|\nabla f(x_k)\|^2 \geq 2\mu\left(f(x_k) - f^\star\right).$$

Assume that $\gamma_k = \alpha_k/(\beta_k L)$. Applying this inequality into the one in Lemma 5 yields

$$f(x_{k+1}) - f^\star \leq \rho\left(f(x_k) - f^\star\right),$$

where $\rho_k = 1 - \Gamma_{\min}/\kappa$ where $\Gamma_{\min} = \min_{k\in[0,T-1]} \alpha_k^2/(2\beta_k)$ and $\kappa = L/\mu$. Suppose that $\rho \in (0, 1)$. Then, by the recursion of the inequality over $k = 0, 1, \ldots, T-1$, we obtain the result. ■

Proof of Theorem 3

We start by assuming that there exists a finite positive constant R such that $\|x_k - x^\star\| \leq R$ where $\{x_k\}$ is generated by (6). This assumption is commonly stated; see e.g., [11]. By the convexity of the objective function f, we have:

$$f(x_k) - f^\star \leq \langle \nabla f(x_k), x_k - x^\star\rangle.$$

By Cauchy-Schwarz's inequality,

$$f(x_k) - f^\star \leq \|\nabla f(x_k)\|\|x_k - x^\star\| \leq R\|\nabla f(x_k)\|.$$

Denote $V_k = f(x_k) - f^\star$. Assume that $\gamma_k = \alpha_k/(\beta_k L)$. Plugging this inequality into the one in Lemma 4, we get

$$V_{k+1} \leq V_k - \Omega_k V_k^2,$$

where $\Omega_k = \alpha_k^2/(2\beta_k L)$. Using this inequality, we have

$$\frac{1}{V_{k+1}} - \frac{1}{V_k} \geq \Omega_{\min} \frac{V_k}{V_{k+1}} \geq \Omega_{\min},$$

where $\Omega_{\min} = \min_{k\in[0,T-1]}\Omega_k$. We reach the last inequality by the fact that $V_{k+1} \leq V_k$ from Lemma 4 with $\gamma = \alpha_k/(\beta_k L)$. By the recursion,

$$\frac{1}{V_T} \geq \frac{1}{V_0} + T\Omega_{\min}.$$

Since $V_0 \geq 0$, the proof is complete. ■

Proof of Theorem 4

Summing the inequality in Lemma 4 with $\gamma_k = \alpha_k/(\beta_k L)$ over $k = 0, 1, \ldots, T-1$ yields

$$\sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq \frac{1}{\Omega_{\min}} \left(f(x_0) - f(x_T)\right),$$

where $\Omega_{\min} = \min_{k\in[0,T-1]}\Omega_k$ and $\Omega_k = \alpha_k^2/(2\beta_k L)$. Using the fact that $\min_{k\in[0,T-1]} \|\nabla f(x_k)\|^2 \leq (1/T)\sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2$, we complete the proof. ■

Proof of Lemma 6

Lemma 6: Consider the optimization problem (1) under Assumption 1. Suppose that $\gamma_k \leq 2/L$. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by (6) with the $K-$greedy quantizer satisfy

$$f(x_{k+1}) \leq f(x_k) - \left(\gamma_k - \frac{L\gamma_k^2}{2}\right)\frac{K}{n}\|\nabla f(x_k)\|^2.$$

Proof: From the definition of the Lipschitz continuity of ∇f and (6), and by the fact $\langle g, Q_G^K(g)\rangle = \|Q_G^K(g)\|^2$,

$$f(x_{k+1}) \leq f(x_k) - \left(\gamma_k - \frac{L\gamma_k^2}{2}\right)\|Q_G^K(\nabla f(x_k))\|^2.$$

Assume that $\gamma_k \leq 2/L$. By the fact that $\|Q_G^K(\nabla f(x_k))\|^2 \geq (K/n)\|\nabla f(x_k)\|^2$, we have the result. ■

REFERENCES

[1] Balcan, Maria Florina, Avrim Blum, Shai Fine, and Yishay Mansour. "Distributed learning, communication complexity and privacy." In *Conference on Learning Theory*, pp. 26-1. 2012.

[2] Tsitsiklis, John N., and Zhi-Quan Luo. "Communication complexity of convex optimization." *Journal of Complexity* 3.3 (1987): 231-243.

[3] Arjevani, Yossi, and Ohad Shamir. "Communication complexity of distributed convex learning and optimization." In *Advances in neural information processing systems*, pp. 1756-1764. 2015.

[4] Bellet, Aurélien, Yingyu Liang, Alireza Bagheri Garakani, Maria-Florina Balcan, and Fei Sha. "A distributed frank-wolfe algorithm for communication-efficient sparse learning." In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 478-486. Society for Industrial and Applied Mathematics, 2015.

[5] Ben-Tal, Ahron, and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications.* Vol. 2. Siam, 2001.

[6] Rabbat, Michael G., and Robert D. Nowak. "Quantized incremental algorithms for distributed optimization." IEEE Journal on Selected Areas in Communications 23.4 (2005): 798-808.

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

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

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

[10] Nutini, Julie, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. "Coordinate descent converges faster with the Gauss-Southwell rule than random selection." In *International Conference on Machine Learning*, pp. 1632-1641. 2015.

[11] Nesterov, Yu. "Efficiency of coordinate descent methods on huge-scale optimization problems." *SIAM Journal on Optimization* 22.2 (2012): 341-362.

[12] 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.*

[13] 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.

[14] 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.