Innovation Compression for Communication-Efficient Distributed Optimization With Linear Convergence
Jiaqi Zhang, Keyou You, Senior Member, IEEE, and Lihua Xie, Fellow, IEEE

Abstract—Information compression is essential to reduce communication cost in distributed optimization over peer-to-peer networks. This article proposes a communication-efficient linearly convergent distributed (COLD) algorithm to solve strongly convex optimization problems. By compressing innovation vectors, which are the differences between decision vectors and their estimates, COLD achieves linear convergence for a class of \( \delta \)-contracted compressors, and we explicitly quantify how the compression affects the convergence rate. Interestingly, our results strictly improve existing results for the quantized consensus problem. Numerical experiments demonstrate the advantages of COLD under different compressors.

Index Terms—Compression, distributed optimization, innovation, linear convergence.

I. INTRODUCTION

We consider the following distributed optimization problem over a network of \( n \) interconnected nodes

\[
\min_{x \in \mathbb{R}^d} f(x) \triangleq \sum_{i=1}^{n} f_i(x)
\]

where the local function \( f_i \) is only known to node \( i \). Each node aims to find an optimal solution \( x^* \in \text{argmin} f(x) \) by communicating with only a subset of nodes that are defined as its neighbors. Such a distributed model has been shown to achieve promising results in many applications [1], [2], [3].

Till now, many novel algorithms have been proposed by transmitting messages with infinite precision, such as the distributed gradient descent algorithm (DGD) [1], the exact first-order algorithm (EXTRA) [4], the network independent step-size algorithm (NIDS) [5], and [6], [7]. As nodes iteratively communicate with neighbors, communication cost can be a bottleneck for the efficiency of distributed optimization. To resolve it, a compressor \( Q(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is unavoidable for each node to send compressed messages, which can be usually encoded with a moderate number of bits. Since \( Q(\cdot) \) is typically highly nonlinear and nonsmooth, how to design a provably convergent distributed algorithm with efficient communication compressions has attracted an increasing attention. To this end, quantized variants of the DGD [1] have been proposed for convex problems [8], [9], [10], [11], [12] and nonconvex problems [13], [14], [15], respectively. Unfortunately, their convergence rates are constrained by the sublinear rate of DGD. Although linear convergence has been achieved in [16], [17], [18], and [19], they are restricted to a specified compressor, which is then extended to a class of stochastic compressors in [20], [21], and [22]. However, their stochastic compressors are assumed to be unbiased in the sense that \( E[Q(x)] = x \). Clearly, this condition excludes deterministic compressors. In fact, the unavoidable round-off error in finite-precision implementation can be regarded as the consequence of a biased compressor.

In this article, we propose a communication-efficient linearly-convergent distributed (COLD) algorithm. The key idea is to compress the innovation—the difference between a state (which can be a decision vector, a gradient, or other variable) and its estimate. Since innovation is expected to eventually decrease to zero, innovation compression is more efficient than directly compressing the decision vectors. Our contributions can be summarized as follows.

1) COLD is the first distributed algorithm with provable linear convergence for \( \mu \)-strongly convex and \( L \)-smooth functions under both unbiased and biased \( \delta \)-contracted compressors.

2) The effect of compression on the convergence rate is explicitly revealed, e.g., the number of iterations to find an \( \epsilon \)-optimal solution is \( O(\max\{\frac{L}{\mu}, \frac{1}{\delta}\}) \log \frac{1}{\epsilon} \) for unbiased and \( \delta \)-contracted compressors, where \( \rho < 1 \) is the spectral gap of the network and \( \delta < 1 \) characterizes the compression resolution. Particularly, \( \delta = 0 \) corresponds to the uncompressed case and COLD recovers the rate of NIDS [5]. With multistep compressions, this rate is improved to \( O(\delta \max\{\frac{L}{\mu}, \frac{1}{\delta}\}) \log \frac{1}{\epsilon} \) for any \( \delta > 0 \).

3) For the consensus problem, i.e., \( f_i(x) = \|x - x_i\|^2 \) in (1), we strictly improve the rate of the state-of-the-art CHOCO-GOSSIP [2] from \( O(\frac{1}{\delta^{(j)}}) \log \frac{1}{\epsilon} \) to \( O(\frac{1}{\delta^{(j)}}) \log \frac{1}{\epsilon} \), and matches the best rate in [23].

The rest of this article is organized as follows. Section II formalizes the problem and introduces the compressors. Section III presents the improved result for the CHOCO-GOSSIP [2], develops the COLD, and provides its theoretical results. Section IV validates the theoretical finding via numerical experiments. Finally, Section V concludes this article. A preliminary conference version has been accepted to IEEE CDC [24], which omits the contents on multistep compressions and all proofs.

Notation: We use \( \nabla f \) to denote the gradient of \( f \). \([A]_{ij}\) denotes the \((i, j)\)th element of \( A \). \( \langle A \rangle \) denotes the \( j \)th largest eigenvalue of \( A \). \( X \cdot Y \) denotes the matrix product. \( \|A\|_F^2 \triangleq \langle A, A \rangle_M \), and \( \|A\|_F^2 \triangleq \|A\|^2 \), where \( M \) is a symmetric matrix, and \( I \) is the identity matrix. 1 denotes a vector with all ones. \( O(\cdot) \) denotes the big-O notation.
II. PROBLEM FORMULATION

A. Compressors

A compressor \( Q(\cdot) : \mathbb{R}^d \to \mathbb{R}^d \) is a (possibly stochastic) mapping for, e.g., quantization or sparsification, and its output can be usually encoded with much fewer bits than its input. This work considers the following class of compressors.

Assumption 1: For some \( \delta > 0 \), \( Q \) satisfies that
\[
E[|Q(x) - x|^2] \leq \delta |x|^2, \quad \forall x \in \mathbb{R}^d
\]
where \( E[\cdot] \) denotes the expectation over \( Q \).

Assumption 1 requires the mean square of the relative compression error to be bounded, which is called \( \delta \)-contracted compressors in [2], [13], [15], and [21]. A compressor \( Q \) is unbiased if \( E[Q(x)] = x \) for all \( x \in \mathbb{R}^d \). Although unbiasedness is essential to the linear convergence in [21], [22], and [20], the widely-used deterministic quantizers are biased. Round-off errors are inevitable in a finite-precision processor and result in a biased compressor, which satisfies Assumption 1 when \( |x| > 2^{-126} \) (IEEE 754 standard [25]). In this view, biased compressors have more practical significance. Some commonly used compressors are given as follows.

Example 1 (Compressors): 

a) Unbiased stochastic quantization: Let \( \text{sgn}(\cdot) \) and \( |\cdot| \) be the element-wise sign function and absolute function, respectively. The compressor is given by \( Q_n(x) = \frac{|x|}{\|x\|} \cdot \text{sgn}(x) + \xi \), where \( \xi \in \mathbb{R}^d \) is a random vector uniformly sampled from \([0,1]^d\). This compressor is unbiased and satisfies Assumption 1 with \( \delta = 2d/4u^2 \) [21].

A common choice is \( u = 2^{-i} \), where each coordinate can be encoded with \( i \) bits. Transmitting \( Q_n(x) \) needs \( (i + 1)d + \text{bits} \) if a scalar can be transmitted with \( b \) bits.

b) Biased quantization: A proper scaling can reduce \( \delta \) but introduce bias. For instance, let \( Q_n(x) \) be defined as in Example 1(a), then \( Q_n(x) = Q_n(x)/\phi \) with \( \phi > 1 \) and \( \xi = 0.5\phi \) is biased. It satisfies Assumption 1 with \( \delta = 1 - \phi^{-1} \) for \( \phi = 1 + d/(4u^2) \) [2].

c) Sparsification: Randomly selecting \( l \leq d \) coordinates of \( x \) (Rand-k) or selecting the random \( l \) coordinates in magnitude (Top-l) gives a biased compressor, which meets Assumption 1 with \( \delta = 1 - 1/d \) [26].

B. Cost Functions and the Communication Network

We focus on strongly convex problems of (1) in this work.

Assumption 2: Each local cost function \( f_i \) is \( \mu \)-strongly convex and \( L \)-Lipschitz smooth, where \( 0 < \mu \leq L \). That is, for all \( x, y \in \mathbb{R}^d, i \in V \)
\[
f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2
\]
\[
\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.
\]

Under Assumption 2, Problem (1) has a unique optimal solution \( x^* = \text{argmin}\, f_i(x) / [27] \).

The interaction between nodes is modeled as an undirected network \( G(\mathcal{V}, \mathcal{E}, W) \), where \( \mathcal{V} = \{1, 2, \ldots, n\} \) is the set of nodes, \( \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \) is the set of edges, and edge \((i,j) \in \mathcal{E}\) if and only if nodes \( i \) and \( j \) can directly exchange compressed messages with each other.

\( G \) is called connected if there exists a path between any pair of nodes. We denote the set of nodes that directly communicate with node \( i \) as the neighbors of \( i \), i.e., \( N_i = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\} \cup \{i\} \). \( W \in \mathbb{R}^{n \times n} \) is an adjacency matrix of \( G \), i.e., \( W_{ij} = 0 \) if \((i,j) \notin \mathcal{E}\). It captures the interactions between neighboring nodes, e.g., \( WX \) can be evaluated via local communications, where \( X = [x_1, \ldots, x_n]^T \) and \( x_i \) denotes the local copy of \( x \) at node \( i \). We make the following standard assumptions [4], [5].

Assumption 3: \( G \) is connected, and \( W \) satisfies that
a) (Symmetry) \( W = W^T \);

b) (Consensus property) null(\( I-W \)) = span(\( 1_n \));

c) (Spectral property) \( -I < W < I \).

An eligible \( W \) can be constructed by the Metropolis rule [4], [23]. Assumption 3 implies that \(-1 < \lambda_0(W) \leq \lambda_2(W) < 1\). We define the spectral gap \( \rho = 1 - \lambda_2(W) \in [0,1) \) to characterize the effect of \( W \) on the convergence rate. Under Assumption 3, Problem (1) is equivalent to the following optimization problem:
\[
\min_{x \in \mathbb{R}^d} F(X) = \sum_{i=1}^{n} f_i(x_i) - (I-W)X = 0. \tag{2}
\]

For simplicity, let \( \tilde{\lambda}_0 = 1 - \lambda_0(W) \in (0,2) \), \( \nabla F(X) = [\nabla f_1(x_1), \ldots, \nabla f_n(x_n)]^T \in \mathbb{R}^{n \times d} \), and we slightly abuse the notation by letting \( Q(X) = [Q(x_1), \ldots, Q(x_n)]^T \in \mathbb{R}^{n \times d} \) where \( X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d} \). Then, \( E[\|Q(X) - X\|^2] \leq \delta \|X\|^2 \) if \( G \) satisfies Assumption 1.

III. COMMUNICATION-EFFICIENT COLD

We propose COLD and explicitly derive its linear convergence rate in this section. The key idea of COLD lies in compressing an innovation vector, which is the discrepancy between the true value of a decision vector and its estimate. Innovation is an important concept in the Kalman filtering theory where it represents the discrepancy between the measurements vector and their optimal predictor. Since the innovation is expected to asymptotically decrease to 0, the compression error vanishes eventually, which is key to the development of COLD. We elaborate this idea by first revisiting the distributed consensus problem and provide a strictly improved rate over CHOCO-GOSSIP [2].

A. Compressed Consensus With Improved Convergence Rate

Distributed consensus is an important special case of Problem (1) with \( f_i(x) = \|x - x_i\|^2, \forall i \in \mathcal{V} \), where nodes are supposed to converge to the average \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). A celebrated algorithm to solve it is proposed in [23] with the update rule \( x_i^{k+1} = \frac{\sum_{j \in N_i}[W_{ij}]x_j^{k+1}}{\sum_{j \in N_i}[W_{ij}]} + \gamma x_i^k, \forall i \in \mathcal{V} \) and \( x_i^0 = x_i \). Let \( X^k = [x_1^k, \ldots, x_n^k]^T \). Its compact form can be given as
\[
X^{k+1} = W X^k \tag{3}
\]

Under Assumption 3, it follows from [23] that \( x^* \) of (3) converges to \( x \) at a linear rate \( O\left(\frac{1}{\sqrt{\log 2}}\right) \) to find an \( \epsilon \)-solution.

However, it requires transmitting \( x^* \) with infinite precision communication. To reduce communication cost by only transmitting \( Q(x^*) \), Aysal et al. [28] proposed \( X^{k+1} = WQ(X^k) \), which was then improved in [8] as \( X^{k+1} = X^k + \gamma (W - I)X^k \). Since \( X^k \) does not converge to zero, we cannot expect the compression error \( Q(x^*) - X^k \) to vanish. Thus, both algorithms cannot converge to \( \bar{x} \).

The first linearly convergent algorithm with a \( \delta \)-contracted compressor is proposed in [2] as follows:
\[
\hat{X}^{k+1} = \hat{X}^k + \gamma (W - I)\hat{X}^k \tag{4}
\]
where \( \gamma \) is a tunable stepsize and \( x_i^0 = x_i, \hat{x}_i^0 = 0 \). We owe the success of (4) to the compression on the innovation \( X^k - \hat{X}^k \), where \( \hat{X}^k \) can be viewed as an estimate of \( X^k \), and reduces to \( X^k \) if the compression error is zero. Since the fixed points of \( x^k \) and \( x^k \) in (4) are both \( \bar{x} \), the innovation as well as its compression error is expected to vanish if nodes have achieved consensus. In this view, we obtain a strictly better...
convergence rate for (4) than that of [2] in the following theorem, the proof of which is relegated to Appendix A.

**Theorem 1:** Let Assumption 1 and 3 hold, and assume $\delta < 1$. Let $\{X^k\}$ and $\{\hat{X}^k\}$ be generated by (4), $\gamma \in (0, \frac{1 - \delta}{(1 + 2\delta)\mu})$, and $e^k = E(q\|X^k - 1X^k\|^2 + \|\hat{X}^{k+1} - X^k\|^2)$, where $q = \frac{1}{(1 + 2\delta)\mu} > 0$. Then, it holds that $e^{k+1} \leq \sigma e^k$, where $\sigma = \max\left\{1 - 2\alpha \frac{1 - \delta}{1 + 2\delta} \gamma \frac{1}{\mu}, \frac{1}{(1 + 2\delta)\mu} \right\} < 1$. In particular, if $\gamma = \frac{1 - \delta}{(3 + \delta)\mu}$, then $\sigma = 1 - \frac{2\alpha}{(1 + \delta)\mu}$.

**Remark 1:** Theorem 1 shows that each node in (4) exactly converges to the average at the linear rate $O\left(\frac{1}{(1 - \delta)\mu} \log \frac{1}{\epsilon}\right)$, which strictly improves the rate $O\left(\frac{1}{\log \frac{1}{\epsilon}}\right)$ in [2], and recovers the best rate of (3) in terms of $\mu$ [23]. Moreover, Theorem 1 shows the linear convergence of both innovations and consensus errors to zero, which is consistent with our observations. An extension of (4) to directed networks is studied in [15].

**B. Development of COLD**

An extension of (4) for the distributed optimization problem (1) is proposed in [2] by directly adding a gradient step in the update of $X^{k+1}$. The resulting algorithm is a quantized version of DGD and can only converge sublinearly.

To achieve a faster convergence rate, we are motivated by the state-of-the-art NIDS [5] that has a linear convergence rate $O(\max\{\frac{1}{\mu}, \frac{1}{\mu}\}|\log \frac{1}{\epsilon}|)$ to find an $\epsilon$-optimal solution of (1), i.e., $\|X_i - x^\star\|^2 \leq \epsilon, \forall i \in \mathcal{N}$. NIDS can be written compactly as:

$$X^{k+1} = \tilde{W}(2X^k - \tilde{X}^{k-1} - \gamma \nabla F(X^k) + \gamma \nabla F(X^{k-1}))$$

where $\tilde{W} = \frac{1}{2}(I + W)$ and the first iteration is initialized as $X^1 = X^0 - \gamma \nabla F(X^0)$. Similar update rules have also been proposed in [6] and [7] for stochastic gradients or left-column weighting matrices. Clearly, it requires transmitting the exact $X^k$ with infinite precision. Similar to the case of the consensus algorithm in (3), directly compressing $X^k$ fails to ensure exact convergence.

We adopt the idea of compressing the innovation vector to design a communication-efficient variant of (5). Instead of compressing the innovation $X^k - \tilde{X}^{k-1}$ in (4), we introduce a new vector $\tilde{X}^k$ and compress its innovation to design the COLD. Specifically, $\tilde{X}^k = 2X^k - X^{k-1} - \gamma \nabla F(X^k) + \gamma \nabla F(X^{k-1})$. Then, (5) can be rewritten as $X^{k+1} = \tilde{W}Y^{k-1}$. This is different from (3) in that each node updates its state as a weighted average of $Y^k$ rather than $X^k$ for all $j \in \mathcal{N}_i$. We can show that $\tilde{X}^k$ converges to $X^k$ by checking the fixed points of (5). Thus, it is reasonable to view $\tilde{X}^k$ as an approximation of $X^k$. By replacing $X^k$ in (4) with $\tilde{X}^k$ and introducing one more stepsize $\tau$, we obtain COLD in the following form:

$$Y^k = 2X^k - \tilde{X}^{k-1} - \gamma \nabla F(X^k) + \gamma \nabla F(X^{k-1})$$

$$\tilde{X}^{k+1} = \tilde{Y}^k + Q(Y^k - \tilde{Y}^k)$$

$$X^{k+1} = X^k + \tau(W - I)\tilde{X}^{k+1}$$

(6)

where $\tilde{Y}^k$ is an auxiliary vector to track $\tilde{X}^k$. The iteration is initialized by $X^1 = Y^0 = X^0 - \gamma \nabla F(X^0)$ and $\tilde{Y}^1 = Q(Y^0)$. Note that the introduction of the tunable stepsize $\tau$ is critical in both theoretical analysis and practical performance. Intuitively, $\gamma$ acts like the stepsize in the standard gradient method and $\tau$ plays a similar role of the stepsize in (4).

**Algorithm 1:** The COLD — From the View of Node $i$.

**Input:** The initial point $x^i_1$. Set $x^i_1 = x^i_0 - \gamma \nabla f_i(x^i_0)$ and $v^i_1 = \tilde{y}^i_1 = y^i_1 = 0$.

1: for $k = 1, 2, \ldots$ do
2: Compute $y^i_k = x^i_k - \gamma \nabla f_i(x^i_k) - \gamma y^i_k$, $q^i_k = Q(y^i_k - \tilde{y}^i_k)$ and $\tilde{y}^i_{k+1} = \tilde{y}^i_k + q^i_k$.
3: Send $q^i_k$ to all neighbors and receive $q^j_k$ from each neighbor $j \in \mathcal{N}_i$.
4: Update

$$\tilde{y}^i_{k+1} = \tilde{y}^i_k + \tau (q^i_k - \sum_{j \in \mathcal{N}_i} [W_{ij}]q^j_k)$$

$$\psi^i_{k+1} = \psi^i_k + \tilde{y}^i_{k+1}$$

$$x^i_{k+1} = x^i_k - \gamma \nabla f_i(x^i_k) - \gamma \psi^i_{k+1}$$

5: end for

By introducing another auxiliary vector $v^i_k \in \mathbb{R}^d$ for each node and defining $\Psi^i = [v^i_1, \ldots, v^i_\tau] \in \mathbb{R}^{n \times \tau}$, COLD in (6) is equivalent to the following form:

$$Y^k = X^k - \gamma \nabla F(X^k) - \gamma \Psi^k$$

$$\tilde{Y}^{k+1} = \tilde{Y}^k + Q(Y^k - \tilde{Y}^k)$$

$$\Psi^{k+1} = \Psi^k + \tau(I - W)\tilde{Y}^{k+1}$$

$$X^{k+1} = X^k - \gamma \nabla F(X^k) - \gamma \Psi^{k+1}$$

(7a)

(7b)

(7c)

(7d)

and $\Psi^1 = \psi^0 = 0$ for $k = 0$. The equivalence can be readily checked by eliminating $\Psi^k$ in (7). COLD of this form is more desirable in implementation since it only requires node $i$ to store one auxiliary variable $v^i_k$ rather than two variables $x^i_{k-1}$ and $\nabla f_i(x^i_{k-1})$. The implementation details are summarized in Algorithm 1, where each node only transmits the compressed vector $Q(Y^i_k - \tilde{Y}^i_k)$ to its neighbors.

**C. Convergence Result for Unbiased Compressors**

The following lemma shows the equivalence between fixed points of (7) and optimal solutions of (2).

**Lemma 1:**

(a) $(X^*, \Psi^*, \tilde{Y}^*)$ is a fixed-point of (7) if and only if $(I - W)X^* = 0, \tilde{Y}^* = X^*, \Psi^* = -\nabla F(X^*)$.

(b) $X^*$ is an optimal solution of (2) if and only if $(I - W)X^* = 0$ and $\nabla F(X^*) \in \text{range}(I - W)$.

**Proof:** (a) It can be easily checked that $(X^*, \Psi^*, \tilde{Y}^*)$ satisfying the condition is a fixed-point. Conversely, it follows from (7d) that $\Psi^* = -\nabla F(X^*)$, and hence $\Psi^* = X^* - \tilde{Y}^* = X^*$. Then, we obtain $(I - W)X^* = 0$ from (7c). (b) It directly follows from the Karush–Kuhn–Tucker theorem for (2).

Since $\Psi^1 = 0$, it follows from (7c) that $\Psi^i \in \text{range}(I - W)$, which implies that $\Psi^i \in \text{range}(I - W)$. Thus, Lemma 1 allows us to focus only on the convergence of (7) to its fixed points. Define $\Theta = \tau^{-1}(I - W) - \gamma I$, which is positive semidefinite in the sequel. We first provide the convergence result of COLD for unbiased compressors under Assumption 1.

**Theorem 2:** Suppose Assumption 1 and 3 hold, $Q(\cdot)$ is unbiased and $\delta < 1$. Let $\gamma \in (0, \frac{1}{\mu})$, $\tau \in (0, \frac{1}{\mu \gamma (2(1-k) + 1)\delta})$ in (7) and $e^k = E(\|X^k - X^k\|^2_\Theta + \|\tilde{Y}^k\|^2_\Theta + \|\psi^k - \psi^k\|^2_\Theta) + \|\Psi^k - \Psi^k\|^2_\Theta$.
where \( \{X^k\}, \{\Psi^k\} \) and \( \{\tilde{Y}^k\} \) are generated by (7). Then, we have \( e^{k+1} \leq \sigma e^k \), where \( \sigma = \max \left\{ \frac{1 - 2\mu L}{\mu + L}, \frac{1 - \gamma^2}{\gamma^2} \right\} < 1 \). In particular, if we set \( \gamma = \frac{1}{\sqrt{2}} \) and \( \tau = \frac{\gamma L}{4(1+\gamma^2)} \), then \( \sigma = \max \left\{ \frac{L}{1 + \gamma^2}, \frac{1 - (1 - \gamma^2)^2}{\gamma^2} \right\} \).

Theorem 2 shows that COLD converges at a linear rate \( \mathcal{O}\left( \frac{L}{1 + \gamma^2} \log \frac{2}{\delta} \right) \) [27]. In comparison with the gradient method of convergence rate \( \mathcal{O}\left( \frac{L}{1 + \gamma^2} \log \frac{2}{\delta} \right) \), we can explicitly quantify how the network and compression affect the rates in terms of simple quantities, i.e., \( \delta \) and \( \rho \).

Since the uncompressd NIDS converges at a linear rate \( \mathcal{O}\left( \max \left\{ \frac{L}{\mu + L}, \frac{1}{\gamma^2} \right\} \log \frac{2}{\delta} \right) \) [5], the \( \delta \)-contracted compression only affects the term related to the network. Particularly, COLD converges at the same rate as NIDS over a network with a "modified" adjacency matrix \( W' \) such that \( \lambda_2(W') = 1 - (1 - \delta)^2 \lambda_2(W) \), and reduces to NIDS if \( \delta = 0 \). Note that COLD is communication-efficient since each node only sends the compressed vector \( \tilde{X}^k \).

Remark 2 (Multistep compressions): It is worth noting that the condition \( \delta < 1 \) can be relaxed to a stronger condition \( \delta < 0 \) at the expense of a higher communication complexity per iteration. Specifically, for an unbiased \( Q(\cdot) \), define a multi-step compression output \( \tilde{Q}(x) = \frac{1}{c} \sum_{i=1}^{c} Q_i(x) \) for some integer \( c > \delta \) where \( Q_i(\cdot) \) is a realization of the random compressor \( Q(\cdot) \). Then, the compression error of \( \tilde{Q} \) is \( \Delta = \delta / c < 1 \). By setting \( c = \lceil 2\delta + 1 \rceil \), Theorem 2 implies that only \( \mathcal{O}(\max \{ \frac{L}{\mu + L}, \frac{1}{\gamma^2} \} \log \frac{2}{\delta} ) \) times of gradient evaluations and \( \mathcal{O}(\delta \max \{ \frac{L}{\mu + L}, \frac{1}{\gamma^2} \} \log \frac{2}{\delta} ) \) times of local communications are needed to find an \( \epsilon \)-optimal solution for any \( \delta > 0 \). Note that the computational complexity is consistent with the NIDS.

Theorem 2 also provides guidance to minimize the overall communication costs for COLD. Consider the unbiased compressor in Example 11. The total number of transmitted bits of node \( n \) to obtain an \( \epsilon \)-optimal solution is \( (l + 1)d + b) \mathcal{O}\left( \max \left\{ \frac{L}{\mu + L}, \frac{1}{\gamma^2} \right\} \log \frac{2}{\delta} \right) \), where \( l + 1 \) is the encoding length. An optimal \( l \) can be obtained by minimizing it, which depends on \( \mu, L, \rho, d, b \), and is not larger than \( \max \{ 0, \log_2 \frac{2}{\delta} \} + 5 \). Thus, transmitting a minimum number of bits (e.g., 1-bit) per iteration may be not optimal for COLD in terms of the total communication load.

The proof of Theorem 2 is based on the following two lemmas. Lemma 2 bounds the innovation and Lemma 3 bounds the distance of decision variables to the fixed points, the proofs of which are relegated to Appendix B. To facilitate presentation, we let \( \hat{X}^k = X^k - \hat{X}^k, \hat{\Psi}^k = \Psi^k - \Psi^*, \tilde{W} = I - \hat{W}, \) and \( \nabla F^{\hat{X}} = \nabla F^{X^k} - \nabla F^\Psi \), where \( \nabla F^{\hat{X}} \) and \( \nabla F^\Psi \) are strong for \( \nabla F^{X^k} \) and \( \nabla F^{X^k} \), respectively.

**Lemma 2:** Under Assumption 1, it holds that

\[
\mathbb{E}\|\hat{X}^{k+1} - X^k\|^2 \leq \frac{2\delta}{1 + \delta} \|\hat{Y}^{k+1} - Y^{k-1}\|^2 + \frac{2\delta^2}{1 + \delta} \|\hat{\Psi}^k + \Psi^k - \Psi^k - \nabla F^\Psi\|^2.
\]

**Lemma 3:** Under Assumption 1 and 3, it holds that

\[
\|\hat{X}^{k+1}\|^2_{\gamma^{-1}} + \mathbb{E}\|\hat{\Psi}^{k+1}\|^2_{\gamma^{-1}} + \|\Psi^{k+1} - \Psi^k\|^2_{\delta} \leq \left( 1 - \frac{2\mu L}{\mu + L} \right) \|\hat{X}^k\|^2_{\gamma^{-1}} + \left( 1 - \frac{\gamma^2}{1 + \delta} \right) \|\nabla F^\Psi\|^2_{\gamma^{-1}} + \|\hat{\Psi}^k\|^2_{\delta} + 2\|\hat{\Psi}^{k+1} - X^{k+1} - Y^k\|^2.
\]

**Proof of Theorem 2:** Notice that \( 2(\hat{Y}^{k+1} + \hat{Y}^{k+1} - Y^k) \) (27) \( 2(\hat{Y}^k + \tau \hat{W} Y^k, \hat{Y}^{k+1} - Y^k) + \|\hat{Y}^{k+1} - Y^k\|^2_{2\tau \hat{W}} \). Since \( Q \) is unbiased and \( \hat{Y}^{k+1} - Y^k = Q(Y^k - \hat{Y}^k) - (Y^k - \hat{Y}^k) \), we have \( \mathbb{E}[2(\hat{Y}^{k+1} + \hat{Y}^{k+1} - Y^k)] = \mathbb{E}[\|\hat{Y}^{k+1} - Y^k\|_{2\tau \hat{W}}^2] \).

Let \( \Lambda = \frac{\tau}{\lambda^2} I \) and add \( \mathbb{E}[\|\hat{Y}^{k+1} - Y^k\|^2] \) to both sides of (9). We obtain that

\[
\|\hat{X}^{k+1}\|^2_{\gamma^{-1}} + \mathbb{E}\|\hat{\Psi}^{k+1}\|^2_{\gamma^{-1}} + \|\Psi^{k+1} - \Psi^k\|^2_{\delta} + \mathbb{E}[\|\hat{Y}^{k+1} - Y^k\|^2_{\Lambda^{-\gamma \hat{W}}} \}
\]

\[
\leq \left( 1 - \frac{2\mu L}{\mu + L} \right) \|\hat{X}^k\|^2_{\gamma^{-1}} + \left( 1 - \frac{\gamma^2}{1 + \delta} \right) \|\nabla F^\Psi\|^2_{\gamma^{-1}} + \frac{2\delta}{1 + \delta} \|\hat{Y}^k - Y^{k-1}\|^2_{\Lambda^{-\gamma \hat{W}}} + \frac{2\delta^2}{1 + \delta} \|\hat{\Psi}^k + \Psi^k - \Psi^k - \nabla F^\Psi\|^2.
\]

Taking expectations on both sides completes the proof. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
D. Convergence Result for Biased Compressors

Now, we turn to general biased compressors. Different from the unbiased case, the convergence requires a small $\delta$ to compensate the biasness caused by the compressor.

Theorem 3: Suppose that Assumption 1 and 3 hold, and $\delta \triangleq 4\sqrt{\gamma}$ < 1. Let $\{X(k), \{\Psi(k)\} \)}$ and $\{Y(k)\}$ be generated by (7), $\gamma \in (0, 2\sqrt{\frac{1}{\mu+L}})$, $\tau \in (0, \frac{1}{\mu L})$, and $e^X = \mathbb{E}||X(k) - X(k)||^2_{\gamma I - 1} + ||\Psi(k)||^2_{\Theta + (1 - \gamma I)Y(k)} + ||\Psi(k) - \Psi(k - 1)||^2_{\Theta + (1 - \gamma I)Y(k - 1)}$. Then, it holds that $e^{k+1} \leq \sigma e^k$, where $\sigma = \max\left\{1 - \frac{2\mu L}{\mu + L}, 1 - \gamma\tau\right\}$.

Compared to the contraction factor $\sigma = \max\left\{1 - \frac{2\mu L}{\mu + L}, 1 - \frac{\gamma\tau}{2}\right\}$ for the unbiased compressors in Theorem 2, Theorem 3 shows that the linear convergence rate for biased compressors has the same dependence on $\gamma, \tau, \mu, L$, and $\rho$ if $\delta$ is small, but reduces faster as $\delta$ increases. Fortunately, a moderate number of encoding bits often leads to a very small $\delta$ since it typically decreases exponentially in the encoding length as shown in Example 11. Moreover, the round-off error caused by finite-precision is a consequence of biased compression with a small $\delta$.

For instance, $\delta < 1.2 \times 10^{-7}$ when using 32-bit floating-point format for $|x| > 2^{-126}$ [25].

Remark 3: Although independently motivated, we found that by setting $\alpha = 1$ in LEAD [21], it reduces to COLD. However, the convergence result in [21] supports only unbiased compressors and does not allow $\alpha = 1$, and thus cannot cover our result. Moreover, the compression rates for biased compressors have the same dependence on $\gamma, \tau, \mu, L$, and $\rho$ if $\delta$ is small, but reduces faster as $\delta$ increases. Fortunately, a moderate number of encoding bits often leads to a very small $\delta$ since it typically decreases exponentially in the encoding length as shown in Example 11. Moreover, the round-off error caused by finite-precision is a consequence of biased compression with a small $\delta$.

IV. NUMERICAL EXPERIMENTS

In this section, we use numerical experiments to validate our theoretical results and compare COLD with existing methods.

Network: We consider $n = 20$ computing nodes connected as an Erdős–Rényi graph, where any two nodes are linked with probability $2m(n) / n$. Then, we use the Metropolis rule [23] to construct $W$ to satisfy Assumption 3. All experiments are repeated 10 times, and we report their average performance.

Tasks: We consider a logistic regression problem on the MNIST dataset [29] with the cost function $f(x_1, \ldots, x_{10}) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{10} l_{ij} \log(u_{ij} / \sum_{j'=1}^{10} u_{ij'}) + \frac{l}{2} \|x_i\|_2^2$, where $m = 60000, l_{ij} \in \{0, 1\}, h_i \in \mathbb{R}^{784}, u_{ij} = \exp(x_i^T h_i)$, and $l = 0.1$. We sort the samples by their labels and then evenly divide them into $n$ parts to create heterogeneous local datasets, and each node has exclusive access to one of them.

Compressors: We test four compressors in Example 1. The unbiased stochastic quantizer with $l = 2$ (C1), the biased quantizer with $l = 2, \phi = 4/3$, and $\xi = 0.544$ (C2), and sparsification compressors which randomly select (C3) or select the largest (C4) 50% coordinates. The required number of bits to transmit their outputs can be computed accordingly. In comparison, nodes in NIDS use 32-bit floating point format. In line with Example 1, the values of $\delta$ of C1–C4 are 49, 36.75, 0.5, 0.5, respectively.

Algorithms: We compare COLD with NIDS [5], LEAD [21], DQOA [22], and CHOCO-SGD [2]. NIDS involves no compression and serves as a baseline. We tune all hyperparameters for these algorithms from the grid $[0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 1, 1.5]$ (2 hyperparameter for CHOCO-SGD, COLD, and NIDS; 3 for LEAD and DQOA). Table I presents the final hyperparameters for C4.

Results: The convergence rates are shown in Figs. 1 and 2. We highlight the following observations.

1) For C1 and C2 satisfying Assumption 1, COLD and LEAD have almost indistinguishable performance from the uncompressed algorithm NIDS w.e.t. iterations [see Fig. 1(a) and (b)]. Moreover, DQOA only converges under C1, and CHOCO-SGD cannot converge to the exact solution.

2) For biased C3 and C4, COLD outperforms other algorithms and has an almost identical convergence rate with NIDS under C4 even
TABLE I  
HYPERPARAMETERS FOUND BY GRID SEARCH FOR C4

| Algorithm | Value           |
|-----------|-----------------|
| COLD      | $\gamma = 1, \tau = 0.3$ |
| NIDS      | $\alpha = 1, c = 0.5$ |
| CHOCO-SGD | $\eta = 0.7, \gamma = 0.5$ |
| LEAD      | $\eta = 1, \gamma = 0.05, \alpha = 0.5$ (suggested by [21]) |
| DQOA      | $\eta = 0.5, \alpha = 0.5, \theta = 0.05$ |

1 As noted in Remark 3, LEAD will have the same performance as COLD if setting $\alpha = 1$ and other parameters appropriately. Nevertheless, we set $\alpha = 0.5$ here to be consistent with [21].

![Graphs](image)

Fig. 1. Optimality gap ($\|f(x^k)\|$) versus number of iterations for different algorithms and compressors. (a) 3-bit unbiased quantization (C1). (b) 3-bit biased quantization (C2). (c) Rand-50% (C3). (d) Top-50% (C4).

![Graphs](image)

Fig. 2. Transmitted bits for algorithm-compressor combinations, which are able to find a solution with $\|f(x^k)\| \leq 10^{-4}$.

V. CONCLUSION

We proposed a communication-efficient distributed algorithm based on innovation compression, which achieves linear convergence even for biased compressors.

APPENDIX A

PROOF OF THEOREM 1

The proof depends on several lemmas. Lemma 4 shows the contraction property of the adjacency matrix $W$.

Lemma 4: For any $W$ satisfying Assumption 3 and $x \in \text{range}\hat{W}$, we have $\lambda_n(W)\|x\|^2 \leq \|\hat{W}x\|^2$.

Proof: By Assumption 3, the eigenvalue 1 is the only eigenvalue of $W$ with magnitude 1, and the corresponding eigenvector is $1_n$. Hence, $\text{range}\hat{W} = \text{range}(I - \frac{1}{n}1_n1_n^T)$. Since $x \in \text{range}\hat{W}$, there exists some $d$ such that $x = (I - \frac{1}{n}1_n1_n^T)d$, which implies $\|x\|_W = d^T(I - \frac{1}{n}1_n1_n^T)W(I - \frac{1}{n}1_n1_n^T)d = d^T(I - \frac{1}{n}1_n1_n^T)(W - \frac{1}{n}1_n1_n^T)(I - \frac{1}{n}1_n1_n^T)d = x^T(W - \frac{1}{n}1_n1_n^T)x$. Note that $\lambda_n(W)1_n \leq W - \frac{1}{n}1_n1_n^T \leq \lambda_2(W)1_n$, and thus the result follows.

Lemma 5: Under Assumption 3 and let $X = 1X^T$. It holds that

$$\|X^k - X\|^2 - \|\hat{X}^k - X^k\|^2 \leq \|X^k - X\|_{I - \gamma\hat{W}}^2 - \|\hat{X}^k - X\|_{I - \gamma\hat{W}}^2 + 2\|X^k - X\|_{I - \gamma\hat{W}}\|X^k - X\|_{I - \gamma\hat{W}}$$

Proof: It follows from (4) and Assumption 3 that $X^k - X = X^k - \hat{X} + \hat{X} - X = X^k - \hat{X} + \gamma\hat{W}(\hat{X} - X)$, which implies that

$$\|X^k - X\|^2 = \|X^k - X\|_{I - \gamma\hat{W}}^2 + \|X^k - \hat{X}\|_{I - \gamma\hat{W}}^2 + 2\|X^k - X\|_{I - \gamma\hat{W}}\|X^k - \hat{X}\|_{I - \gamma\hat{W}}$$

Since $2(a, b)_M = \|a + b\|_M^2 - \|a\|_M^2 - \|b\|_M^2$ for any symmetric matrix $M$, we have $-2\|X^k - X\|_{I - \gamma\hat{W}}\|X^k - \hat{X}\|_{I - \gamma\hat{W}} = \|X^k - \hat{X}\|_{I - \gamma\hat{W}}^2 - \|X^k - \hat{X}\|_{I - \gamma\hat{W}}^2 = \|X^k - X\|_{I - \gamma\hat{W}}^2 - \|X^k - X\|_{I - \gamma\hat{W}}^2 - \|X^k - X\|_{I - \gamma\hat{W}}^2$, which jointly with (13) completes the proof.

Lemma 6: Under Assumption 1 and 3, it holds that

$$\mathbb{E}\|\hat{X}^{k+1} - X\|^2 \leq \delta \left(\|\hat{X}^k - X^k\|^2_{I + \gamma\hat{W}} - \|X^k - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}}\right).$$

Proof: We have from (4) that

$$\mathbb{E}\|\hat{X}^{k+1} - X\|^2 = \mathbb{E}\|Q(X^k - \hat{X}) - (X^k - \hat{X})\|^2$$

$$\leq \delta \|\hat{X}^k - X\|^2 + \delta \|\hat{X}^k - X^{k-1}\|^2 + \gamma\hat{W}(\hat{X}^k - X\|^2$$

$$= \delta \left(\|\hat{X}^k - X^{k-1}\|^2_{I + \gamma\hat{W}} - \|X^k - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}}\right)$$

where the inequality follows from Assumption 1, and the last equality follows from $-2\|X^k - X\|_{I + \gamma\hat{W}}\|X^k - \hat{X}\|_{I + \gamma\hat{W}} = -\|X^k - X\|_{I + \gamma\hat{W}}^2 + \|X^k - \hat{X}\|_{I + \gamma\hat{W}}^2$. Combining the abovementioned completes the proof.

Proof of Theorem 1: Let $q = \frac{1 + \gamma\hat{W}}{1 - \gamma\hat{W}} \delta > \delta$. Multiplying (12) by $q$ and then adding it to (14), we obtain

$$\|X^k - X\|_{I + \gamma\hat{W}}^2 + \mathbb{E}\|\hat{X}^{k+1} - X\|^2 \leq \|X^k - X\|_{I + \gamma\hat{W}}^2 - \|X^k - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} - \|\hat{X}^k - X\|^2_{I + \gamma\hat{W}}$$

$$\leq \|X^k - X\|_{I + \gamma\hat{W}}^2 + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} - \|\hat{X}^k - X\|^2_{I + \gamma\hat{W}},$$

$$\leq \|X^k - X\|_{I + \gamma\hat{W}}^2 + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} + \|\hat{X}^{k-1} - X^{k-1}\|^2_{I + \gamma\hat{W}} - \|\hat{X}^k - X\|^2_{I + \gamma\hat{W}}.$$
For any \( \gamma \in (0, \lambda_n^{-1}) \), it follows from Assumption 3 that \( \tilde{W} = \sqrt{\gamma W} \geq 0 \). We have \( -\gamma W(qI - \delta I - (q - \delta) \gamma \tilde{W}) = \frac{2\gamma W(W - \gamma_n I)W}{1 - \gamma_n} \leq 0 \). It follows from (15) that
\[
\|X^k - \tilde{X}\|^2_{\tilde{W}} + \mathbb{E}[\|	ilde{X}^k - X^k\|^2_{\tilde{W}}] \\
\leq \|X^{k+1} - \tilde{X}\|^2_{\tilde{W}} + \mathbb{E}[\|	ilde{X}^k - X^k\|^2_{\tilde{W}}] \\
\leq \|X^{k-1} - \tilde{X}\|^2_{\tilde{W}} + \mathbb{E}[\|	ilde{X}^k - X^{k-1}\|^2_{\tilde{W}}] \\
\leq \sigma^2 \left( \|X^{k-1} - \tilde{X}\|^2_{\tilde{W}} + \|\tilde{X}^k - X^{k-1}\|^2_{\tilde{W}} \right)
\]
where the second inequality follows from Lemma 4 by noticing that each column of \( X^{k-1} - \tilde{X} = (I - \frac{1}{\eta} I)^T \tilde{X} \) belongs to range \( \tilde{W} \), and \( \sigma \) is defined as follows:
\[
\sigma = \max \left\{ 1 - \left( 1 + \frac{1}{\eta} \right) \gamma \rho, \delta + \frac{1}{\eta} \gamma \lambda_n \right\} \\
= \max \left\{ 1 - \frac{2\gamma \rho}{1 + \gamma \lambda_n}, \frac{\delta \gamma \lambda_n}{1 - \gamma \lambda_n} \right\} < 1
\]
where the last inequality follows from \( \gamma < \frac{1 - \delta}{\eta (1 + \delta) \lambda_n} \). Then, the first part of Theorem 1 follows by taking full expectations on both sides. Substituting \( \gamma = \frac{1 - \delta}{\eta (1 + \delta) \lambda_n} \) into (16), we obtain \( \sigma = 1 - \frac{(1 - \delta) \rho}{2\eta \lambda_n} \). The desired result follows.

**Proof of Lemma 2**

We have
\[
Y^{k} \overset{(7a)}{=} X^{k} - \gamma \Psi^{k} - \gamma \nabla F^{k} \\
Y^{k-1} + X^{k} - X^{k-1} + \gamma (\Psi^{k-1} - \Psi^{k} + \nabla F^{k-1} - \nabla F^{k}) \\
Y^{k-1} + \gamma (\Psi^{k-1} - 2 \Psi^{k}) - \gamma \nabla F^{k}
\]
Lemma 1 holds if \( Y^{k-1} + \gamma (\Psi^{k-1} - \Psi^{k}) - \gamma \tilde{W}\Psi^{k} - \gamma \nabla F^{k} \).

Therefore
\[
\mathbb{E}[\|\tilde{W}^{k+1} - Y^{k}\|^2] \\
\overset{(7b)}{=} \mathbb{E}[\|Q(Y^{k} - \tilde{W}) - (Y^{k} - \tilde{W})\|^2] \leq \delta \|\tilde{W}^{k} - Y^{k}\|^2 \\
\overset{(7c)}{=} \delta \|\tilde{W}^{k} - Y^{k-1} + \gamma \tilde{W}\Psi^{k} + \gamma (\Psi^{k-1} - \Psi^{k}) + \gamma \nabla F^{k}\|^2 \\
\leq \frac{2\delta}{1 + \delta} \|\tilde{W}^{k} - Y^{k-1}\|^2 + \frac{2\delta \gamma^2}{1 + \delta} \|	ilde{W}^{k} + \Psi^{k} - \Psi^{k-1} + \nabla F^{k}\|^2
\]
where the first inequality follows from Assumption 1 and the last inequality used the relation \( \|a + b\|^2 \leq (1 + c)\|a\|^2 + (1 + \frac{1}{c})\|b\|^2 \), \( \forall a, b \) with \( c = \frac{1 - \delta}{1 + \delta} \).

**B. Proof of Lemma 3**

We need several lemmas first. **Lemma 7:** Under Assumption 1 and 3, the following relation holds for all \( k \geq 1 \):
\[
\|X^{k+1}\|^2_{\gamma \tilde{I}} + \|\tilde{W}^{k+1}\|^2_{\tilde{I}} = \|X^{k}\|^2_{\gamma \tilde{I}} + \|	ilde{W}^{k}\|^2_{\tilde{I}}
\]

**Proof of Lemma 3:** For any symmetric \( \Theta \), we have \( \frac{2}{\|a\|^2_{\Theta}} = \|a\|^2_{\Theta} - \|b\|^2_{\Theta}, \forall a, b \), and hence
\[
2\langle X^{k}, \nabla F^{k} \rangle + 2\langle \tilde{W}^{k}, X^{k+1} - X^{k} \rangle
\]
where the inequality follows from Lemma 8 and (7c). Therefore
\[
\|\hat{X}^{k+1}\|_{\gamma^{-1}I}^2 + \|\hat{\Psi}^{k+1}\|_{\Theta}^2 - 2\langle \hat{\Psi}^{k+1}, \hat{X}^{k+1} - Y^k \rangle
\]
\[
+ \|\Psi^{k+1} - \Psi^k\|^2_{\Theta^2}
\]
\[
\overset{(18)}{=} \|\hat{X}^{k}\|_{\gamma^{-1}I}^2 + \|\hat{\Psi}^{k}\|_{\Theta}^2 - \|X^{k+1} - X^k\|_{\gamma^{-1}I}^2
\]
\[
- 2\langle \hat{X}^{k}, \nabla F^{k}\rangle - 2\langle \hat{\Psi}^{k}, X^{k+1} - X^k \rangle
\]
\[
\overset{(21)}{\leq} \left( 1 - \frac{2\mu L\gamma}{\mu + L} \right) \|\hat{X}^{k}\|_{\gamma^{-1}I}^2 - \|\hat{\Psi}^{k+1}\|_{\Theta}^2 + \|\Psi^{k}\|_{\Theta}^2 + \left( 1 - \frac{2\gamma^{-1}}{\mu + L} \right) \|\nabla F^{k}\|_{\Theta}^2.
\]

The desired result is then obtained. ■

REFERENCES

[1] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” IEEE Trans. Autom. Control, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[2] A. Koloskova, S. Stich, and M. Jaggi, “Decentralized stochastic optimization and gossip algorithms with compressed communication,” in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 3478–3487.

[3] J. Zhang, K. You, and K. Cai, “Distributed dual gradient tracking for resource allocation in unbalanced networks,” IEEE Trans. Signal Process., vol. 68, pp. 2186–2198, Mar. 2020.

[4] W. Shi, Q. Ling, G. Wu, and W. Yin, “EXTRA: An exact first-order algorithm for decentralized consensus optimization,” SIAM J. Optim., vol. 25, no. 2, pp. 944–966, 2015.

[5] Z. Li, W. Shi, and M. Yan, “A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates,” IEEE Trans. Signal Process., vol. 67, no. 17, pp. 4494–4506, Sep. 2019.

[6] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, “D^2: Decentralized training over decentralized data,” in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 4848–4856.

[7] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, “Exact diffusion for distributed optimization and learning—Part I: Algorithm development,” IEEE Trans. Signal Process., vol. 67, no. 3, pp. 708–723, Feb. 2019.

[8] R. Carli, F. Fagnani, P. Frasca, and S. Zampieri, “Gossip consensus algorithms via quantized communication,” Automatica, vol. 46, no. 1, pp. 70–80, 2010.

[9] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, “An exact quantized decentralized gradient descent algorithm,” IEEE Trans. Signal Process., vol. 67, no. 19, pp. 4934–4947, Oct. 2019.

[10] H. Li, C. Huang, Z. Wang, G. Chen, and H. G. A. Umar, “Computation-efficient distributed algorithm for convex optimization over time-varying networks with limited bandwidth communication,” IEEE Trans. Signal Inf. Process. Netw., vol. 6, pp. 140–151, Jan. 2020.

[11] Y. Yi and Y. Hong, “Quantized subgradient algorithm and data-rate analysis for distributed optimization,” IEEE Trans. Control Netw. Syst., vol. 1, no. 4, pp. 380–392, Dec. 2014.

[12] T. T. Doan, S. T. Maguluri, and J. Romberg, “Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach,” IEEE Trans. Autom. Control, vol. 66, no. 10, pp. 4469–4484, Oct. 2021.

[13] T. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, “Communication compression for decentralized training,” in Proc. Adv. Neural Inf. Process. Syst., 2018, vol. 31, pp. 7652–7662.

[14] Y. Lu and C. De Sa, “Moniqua: Modulo quantized communication in decentralized SGD,” in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 6415–6425.

[15] H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, “Quantized decentralized stochastic learning over directed graphs,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 9324–9333.

[16] C.-S. Lee, N. Michelusi, and G. Scutari, “Finite rate quantized distributed optimization with geometric convergence,” in Proc. 52nd Asilomar Conf. Signals, Syst., Comput., 2018, pp. 1876–1880.

[17] S. Magrinius, H. Shokri-Ghadikolaei, and N. Li, “On maintaining linear convergence of decentralized learning and optimization under limited communication,” IEEE Trans. Signal Process., vol. 68, pp. 6101–6116, Oct. 2020.

[18] Y. Xiong, L. Wu, K. You, and L. Xie, “Quantized distributed gradient tracking algorithm with linear convergence in directed networks,” IEEE Trans. Autom. Control, early access, Nov. 2022.

[19] Y. Kajiyama, N. Hayashi, and S. Takai, “Linear convergence of consensus-based quantized optimization for smooth and strongly convex cost functions,” IEEE Trans. Autom. Control, vol. 66, no. 3, pp. 1254–1261, Mar. 2021.

[20] Y. Liao, Z. Li, K. Huang, and S. Pu, “A Compressed gradient tracking for decentralized optimization with linear convergence,” IEEE Trans. Autom. Control, vol. 67, no. 10, pp. 5622–5629, Oct. 2022.

[21] X. Liu, Y. Li, R. Wang, J. Tang, and M. Yan, “Linear decentralized optimization with compression,” in Proc. Int. Conf. Learn. Representations, 2021.

[22] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtarik, and S. Stich, “A linearly convergent algorithm for decentralized optimization: Sending less bits for free,” in Proc. Int. Conf. Artif. Intell. Statist., 2021, pp. 4087–4095.

[23] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” Syst. Control Lett., vol. 53, no. 1, pp. 65–78, 2004.

[24] J. Zhang and K. You, “Linearly convergent distributed optimization methods with compressed innovation,” in Proc. IEEE 60th Conf. Decis. Control, 2021, pp. 248–252.

[25] IEEE, IEEE Standard for Floating-Point Arithmetic, IEEE Standard 754–2019, 2019.

[26] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, “SparseSGD with memory,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 4447–4458.

[27] Y. Nesterov, Introductory Lectures on Convex Optimization: A. Basic Course, vol. 87, Berlin, Germany: Springer, 2013.

[28] T. C. Aysal, M. J. Coates, and M. G. Rabbat, “Distributed average consensus with dithered quantization,” IEEE Trans. Signal Process., vol. 56, no. 10, pp. 4905–4918, Oct. 2008.

[29] Y. LeCun, “The MNIST database of handwritten digits,” 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/