Communication Compression for Distributed Nonconvex Optimization

Xinlei Yi , Shengjun Zhang , Tao Yang , *Senior Member, IEEE*, Tianyou Chai , *Life Fellow, IEEE*, and Karl Henrik Johansson , *Fellow, IEEE*

***Abstract*—In this article, we consider distributed non-convex optimization with the cost functions being distributed over agents. Noting that information compression is a key tool to reduce the heavy communication load for distributed algorithms as agents iteratively communicate with neighbors, we propose three distributed primal–dual algorithms with compressed communication. The first two algorithms are applicable to a general class of compressors with bounded relative compression error and the third algorithm is suitable for two general classes of compressors with bounded absolute compression error. We show that the proposed distributed algorithms with compressed communication have comparable convergence properties as state-of-the-art algorithms with exact communication. Specifically, we show that they can find first-order stationary points with sublinear convergence rate $\mathcal{O}(1/T)$ when each local cost function is smooth, where T is the total number of iterations, and find global optima with linear convergence rate under an additional condition that the global cost function satisfies the Polyak–Łojasiewicz condition. Numerical simulations are provided to illustrate the effectiveness of the theoretical results.**

***Index Terms*—Communication compression, distributed optimization, linear convergence, nonconvex optimization, Polyak–Łojasiewicz (P–Ł) condition.**

I. INTRODUCTION

WE CONSIDER distributed nonconvex optimization. Specifically, consider a network of n agents, each of

which has a private local (possibly nonconvex) cost function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$. The whole network aims to solve the following optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x). \qquad (1)$$

Throughout this article, we assume each f_i is smooth. Note that each agent alone cannot solve the above optimization problem since it does not know other agents' local cost functions. Therefore, agents need to communicate with each other through an underlying communication network. Distributed nonconvex optimization has wide applications, such as power allocation in wireless ad hoc networks [1], distributed clustering [2], dictionary learning [3], and empirical risk minimization [4].

The problem (1) has been extensively studied in the literature, e.g., [5], [6], [7], [8], [9], [10], [11], just to name a few. Due to nonconvexity, these studies typically showed that first-order stationary points can be found at a sublinear convergence rate. For example, Hong et al. [5], [6], [7], [8], [9] showed that first-order stationary points can be found with an $\mathcal{O}(1/T)$ convergence rate, where T is the total number of iterations. However, the algorithms proposed in these studies require significant amount of data exchange as agents iteratively communicate with neighbors. Noting that communication bandwidth and power are limited, it is vital to propose communication-efficient distributed algorithms. In this article, we propose distributed algorithms with compressed communication to improve communication efficiency.

A. Related Works and Motivation

It is straightforward to combine existing distributed algorithms and communication compression directly. However, such a simple strategy does not converge to the accurate solution due to the compression error and even may lead to divergence as the compression error would accumulate. Examples have been provided in [12] and [13] to illustrate this. Therefore, communication compression in distributed algorithms has gained considerable attention recently.

When each local cost function is convex, various distributed algorithms with compressed communication have been proposed. For example, Alistarh et al. [14] and Horváth et al. [15] used unbiased compressors with bounded relative compression error to design distributed stochastic gradient descent (SGD) algorithms; Koloskova et al. [16] employed biased but contractive compressors to design a distributed SGD algorithm; Reisizadeh et al. [17] and Liu et al. [18], [19] utilized unbiased compressors to respectively design distributed gradient descent

and primal–dual algorithms; Nedic et al. [20] and Zhu et al. [21] made use of the standard uniform quantizer to respectively design distributed subgradient methods and alternating direction method of multipliers approaches; Yuan et al. [22], [23] and Doan et al. [24], respectively, adopted the unbiased random and adaptive quantization to design distributed projected subgradient algorithms; Yi and Hong [25] and Lee et al. [26], [27], [28], [29] exploited the standard uniform quantizer with dynamic quantization level to respectively design distributed subgradient and primal–dual algorithms; and Lei et al. [30] applied the standard uniform quantizer with a fixed quantization level to design a distributed gradient descent algorithm. The compressors mentioned above can be unified into three general classes. Specifically, Liao et al. [31] proposed a wider class of compressors with bounded relative compression error which covers the compressors used in [13], [14], [15], [16], [17], [18], and [19]; Khirirat et al. [32] considered a general class of compressors with globally bounded absolute compression error, which accommodates the compressors used in [20], [21], [22], [23], and [24]; and Zhang et al. [33] studied a general class of compressors with locally bounded absolute compression error, which contains the compressors used in [12], [25], [26], [27], [28], [29], and [30]. These studies also analyzed the convergence properties of the proposed algorithms. Especially, some of them showed that the achieved convergence rates under compressed communication are comparable to and even match those under exact communication. For instance, linear convergence was achieved in [18], [19], [26], [27], [28], [29], [31], and [33] under the standard strong convexity assumption.

While various algorithms with compressed communication have been designed for distributed convex optimization, communication compression for distributed nonconvex optimization is relatively less studied because the analysis is more challenging due to the nonconvexity. Moreover, when considering distributed nonconvex optimization, most of existing distributed algorithms with compressed communication are SGD algorithms although different types of compressors have been used. For instance, Lu and De Sa [12] used the modular arithmetic for communication quantization (Moniqua); Tang et al. [34] used unbiased compressors with bounded relative or absolute compression error; Koloskova et al. [35], [36], [37], [38] used biased but contractive compressors; and Reisizadeh et al. [39] used unbiased compressors with bounded absolute compression error. These studies also analyzed the convergence properties of the proposed algorithms. For instance, Lu et al. [12], [34], [35], [36], [37], [38] showed that the proposed SGD algorithms with compressed communication achieve linear speedup convergence rate $\mathcal{O}(1/\sqrt{nT})$, which is the same as that achieved by distributed SGD algorithms with exact communication. Observing this, one core theoretical question arises.

(Q1) Under compressed communication, can first-order stationary points be found with the well-known $\mathcal{O}(1/T)$ convergence rate?

On the other hand, noting that it has been shown in [9], [40], and [41] that global optima of nonconvex optimization can be linearly found if the global cost function satisfies the Polyak–Łojasiewicz (P–Ł) condition, another core theoretical question arises.

(Q2) Under compressed communication, can global optima be linearly found when the global cost function satisfies the P–Ł condition?

B. Main Contributions

In this article, we provide positive answers to the above questions. More specifically, the contributions of this article are summarized as follows.

(C1) We first use a general class of compressors with bounded relative compression error, which incorporates various commonly used compressors including unbiased compressors and biased but contractive compressors, to design a communication-efficient distributed algorithm (Algorithm 1). This algorithm only requires each agent to communicate one compressed variable with its neighbors per iteration. We show that this compressed communication algorithm has comparable convergence properties as state-of-the-art algorithms with exact communication. Specifically, we show in Theorem 1 that it can find a first-order stationary point with the well-known $\mathcal{O}(1/T)$ convergence rate; thus, (Q1) is answered. Moreover, if the global cost function satisfies the P–Ł condition, we show in Theorem 2 that it can find a global optimum with linear convergence rate; thus, (Q2) is answered.

(C2) We then propose an error feedback based compressed communication algorithm (Algorithm 2) for biased compressors particularly. This algorithm can correct the bias induced by biased compressors under the cost that it requires each agent to communicate two compressed variables with its neighbors per iteration. We show in Theorems 3 and 4 that this algorithm has similar convergence properties as the first algorithm, which respectively answer (Q1) and (Q2).

(C3) We finally use two general classes of compressors with globally and locally bounded absolute compression error, which cover various commonly used compressors including unbiased compressors with bounded variance, random/adaptive/uniform quantization, and even 1-bit binary quantizer, to design a communication-efficient distributed algorithm (Algorithm 3). This algorithm also only requires each agent to communicate one compressed variable with its neighbors per iteration. When the compressors have globally bounded absolute compression error, we show in Theorems 5 and 6 that this algorithm has similar convergence properties as the first algorithm, which respectively answer (Q1) and (Q2). When the compressors have locally bounded absolute compression error, we show in Theorem 7 that this algorithm can find a global optimum with linear convergence rate if the global cost function satisfies the P–Ł condition and a lower bound on the corresponding P–Ł constant is known a priori, which answers (Q2).

In summary, the main contribution of this article is to propose three distributed primal–dual algorithms with compressed communication for distributed nonconvex optimization, which have comparable convergence properties as state-of-the-art algorithms with exact communication. This is a significant theoretical development, and to the best of our knowledge, it is the first time to achieve this.

C. Outline

The rest of this article is organized as follows. Section II introduces some preliminaries. Section III presents the problem formulation. Sections IV–VI provide three communication-efficient distributed algorithms and analyze their convergence properties. Section VII gives numerical simulations. Finally, Section VIII concludes this article.

Notations: $\mathbb{N}_0$ denotes the set of nonnegative integers. $[n]$ denotes the set $\{1, \ldots, n\}$ for any positive constant integer n. $\|\cdot\|_p$ represents the p-norm for vectors or the induced p-norm for matrices, and the subscript is omitted when $p = 2$. Given a differentiable function f, ∇f denotes its gradient. $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column one (zero) vector of dimension n. $\mathbf{I}_n$ is the n-dimensional identity matrix. $\mathrm{col}(z_1, \ldots, z_k)$ is the concatenated column vector of vectors $z_i \in \mathbb{R}^{d_i}$, $i \in [k]$. Given a vector $[x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, $\mathrm{diag}([x_1, \ldots, x_n])$ is a diagonal matrix with the ith diagonal element being x_i. The notation $A \otimes B$ denotes the Kronecker product of matrices A and B. Given two symmetric matrices M and N, $M \geq N$ means that $M - N$ is positive semidefinite. $\mathrm{null}(A)$ is the null space of matrix A. $\rho(\cdot)$ stands for the spectral radius for matrices and $\rho_2(\cdot)$ indicates the minimum positive eigenvalue for matrices having positive eigenvalues. For any square matrix A, denote $\|x\|_A^2 = x^\top A x$.

II. PRELIMINARIES

In this section, we briefly introduce algebraic graph theory and the P–Ł condition.

A. Algebraic Graph Theory

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ denote a weighted undirected graph with the set of vertices (nodes) $\mathcal{V} = [n]$, the set of links (edges) $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and the weighted adjacency matrix $A = A^\top = (a_{ij})$ with nonnegative elements a_{ij}. A link of $\mathcal{G}$ is denoted by $(i, j) \in \mathcal{E}$ if $a_{ij} > 0$, i.e., if vertices i and j can communicate with each other. It is assumed that $a_{ii} = 0$ for all $i \in [n]$. Let $\mathcal{N}_i = \{j \in [n] : a_{ij} > 0\}$ and $\deg_i = \sum_{j=1}^n a_{ij}$ denote the neighbor set and weighted degree of vertex i, respectively. The degree matrix of graph $\mathcal{G}$ is $\mathrm{Deg} = \mathrm{diag}([\deg_1, \ldots, \deg_n])$. The Laplacian matrix is $L = (L_{ij}) = \mathrm{Deg} - A$. A path of length k between vertices i and j is a subgraph with distinct vertices $i_0 = i, \ldots, i_k = j \in [n]$ and edges $(i_j, i_{j+1}) \in \mathcal{E}, j = 0, \ldots, k-1$. An undirected graph is connected if there exists at least one path between any two distinct vertices.

B. Polyak–Łojasiewicz Condition

Let $f(x) : \mathbb{R}^d \mapsto \mathbb{R}$ be a differentiable function. Let $\mathbb{X}^* = \arg\min_{x \in \mathbb{R}^p} f(x)$ and $f^* = \min_{x \in \mathbb{R}^d} f(x)$. Moreover, we assume that $f^* > -\infty$.

Definition 1: The function f satisfies the P–Ł condition with constant $\nu > 0$ if

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*) \quad \forall x \in \mathbb{R}^d. \tag{2}$$

It is straightforward to see that every (essentially or weakly) strongly convex function satisfies the P–Ł condition. The P–Ł condition implies that every stationary point is a global minimizer. But unlike the (essential or weak) strong convexity, the P–Ł condition alone does not imply convexity of f. Moreover, it does not imply that the global minimizer is unique either. In fact, P–Ł condition generalizes strong convexity to nonconvex functions. The function $f(x) = x^2 + 3\sin^2(x)$ given in [40] is an example of a nonconvex function satisfying the P–Ł condition with $\nu = 1/32$. Moreover, it was shown in [42] that the loss functions in some applications satisfy the P–Ł condition in the local region near a local minimum. Moreover, Fazel et al. [43] proved that the cost function of the policy optimization for the linear quadratic regulator problem is nonconvex and satisfies the P–Ł condition.

III. PROBLEM FORMULATION

In this section, we introduce three general classes of compressors and provide the assumptions on the communication network and cost functions.

A. Compressors

To improve communication efficiency, we consider the scenario that the communication between agents is compressed. Specifically, we consider a class of compressors with bounded relative compression error, and two classes of compressors, respectively, with globally and locally bounded absolute compression error satisfying the following assumptions.

Assumption 1: The compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ satisfies

$$\mathbf{E}_\mathcal{C}\left[\left\|\frac{\mathcal{C}(x)}{r} - x\right\|^2\right] \leq (1 - \varphi)\|x\|^2 \quad \forall x \in \mathbb{R}^d \tag{3}$$

for some constants $\varphi \in (0, 1]$ and $r > 0$. Here, $\mathbf{E}_\mathcal{C}[\cdot]$ denotes the expectation over the internal randomness of the stochastic compression operator $\mathcal{C}$.

From (3), we have

$$\mathbf{E}_\mathcal{C}[\|\mathcal{C}(x) - x\|^2] = \mathbf{E}_\mathcal{C}\left[\left\|r\left(\frac{\mathcal{C}(x)}{r} - x\right) + (r-1)x\right\|^2\right]$$

$$\leq 2r^2 \mathbf{E}_\mathcal{C}\left[\left\|\frac{\mathcal{C}(x)}{r} - x\right\|^2\right] + 2(1 - r)^2\|x\|^2$$

$$\leq r_0\|x\|^2 \quad \forall x \in \mathbb{R}^d \tag{4}$$

where $r_0 = 2r^2(1 - \varphi) + 2(1 - r)^2$. Therefore, the class of compressors satisfying Assumption 1 is the same as that used in [31]. As explained in [31], the class of compressors satisfying Assumption 1 is broad. It incorporates all the unbiased compressors with bounded relative compression error[1] and biased but contractive compressors,[2] such as random quantization and sparsification, which are commonly used in the literature, e.g., [14], [15], [16], [17], [18], [19], [34], [35], [36], [37], [38], [44]. It also includes some biased and noncontractive compressors, such as the norm-sign compressor. Moreover, it is straightforward to check that the class of compressors satisfying Assumption 1 also covers the three classes of biased compressors considered in [13]. In other words, Assumption 1 is weaker than various commonly used assumptions for compressors in the literature.

Assumption 2: The compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ satisfies

$$\mathbf{E}_\mathcal{C}[\|\mathcal{C}(x) - x\|_p^2] \leq C \quad \forall x \in \mathbb{R}^d \tag{5}$$

for some real number $p \geq 1$ and constant $C \geq 0$.

The same class of compressors satisfying Assumption 2 has also been used in [32], which incorporates the deterministic quantization used in [20], [21], [22], and [24] and the unbiased random quantization used in [22], [23], [34], and [39].

Assumption 3: The compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ satisfies

$$\|\mathcal{C}(x) - x\|_p \leq (1 - \varphi) \quad \forall x \in \{x \in \mathbb{R}^d : \|x\|_p \leq 1\} \tag{6}$$

for some real number $p \geq 1$ and constant $\varphi \in (0, 1]$.

The same class of compressors satisfying Assumption 3 has also been used in [33], which covers the standard

---

[1]A compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is unbiased with bounded relative compression error (or just unbiased for simplicity) if for all $x \in \mathbb{R}^d$, $\mathbf{E}_\mathcal{C}[\mathcal{C}(x)] = x$ and there exists a constant $C \geq 0$ such that $\mathbf{E}_\mathcal{C}[\|\mathcal{C}(x) - x\|^2] \leq C\|x\|^2$.

[2]A compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is contractive if there exists a constant $\varphi \in (0, 1]$ such that $\mathbf{E}_\mathcal{C}[\|\mathcal{C}(x) - x\|^2] \leq (1 - \varphi)\|x\|^2$, $\forall x \in \mathbb{R}^d$.

uniform quantizers with dynamic and fixed quantization levels, respectively, used in [25], [26], [27], [28], [29], and [30], and the Moniqua used in [12]. Moreover, as pointed out in [33], the 1-bit binary quantizer satisfies Assumption 3. The difference between Assumptions 2 and 3 is that the former is a global assumption while the latter is a local assumption. It should be pointed out that all Assumptions 1–3 do not require the compressors to be unbiased. Note that the inequalities in Assumptions 1–3 are different, and no one can imply another. Therefore, the three types of compressors are different from each other, and no one type is more restrictive than or can imply another. Moreover, the intersection of each pair of the three types of compressors is nonempty. For example, as explained in Section VII, the norm-sign compressor satisfies both Assumptions 1 and 3. Therefore, the three types of compressors are not mutually exclusive.

The above three general classes of compressors cover most of existing compressors used in machine learning and signal processing applications, which substantiate the generality of the results later in this article.

B. Communication Network and Cost Functions

The following assumptions for the problem (1) are made.

Assumption 4: The underlying communication network is modeled by an undirected and connected graph $\mathcal{G}$.

Assumption 5: The minimum function value of the optimization problem (1) is finite.

Assumption 6: Each local cost function $f_i(x)$ is smooth with constant $L_f > 0$, i.e., it is differentiable and

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_f \|x - y\| \quad \forall x, y \in \mathbb{R}^d. \quad (7)$$

Assumption 7: The global cost function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$.

Assumptions 4–6 are standard in the literature to guarantee the well-known $\mathcal{O}(1/T)$ convergence rate for distributed algorithms finding the first-order stationary points for nonconvex optimization problems. Assumption 7 is weaker than the assumption that the global or each local cost function is strongly convex but it still can guarantee linear convergence. Note that the convexity of the cost functions and the boundedness of their gradients are not assumed. We also make no assumptions on the boundedness of the deviation between the gradients of local cost functions. In other words, we do not assume that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2$ is bounded. Moreover, we do not assume that the optimal set is a singleton or finite set either.

IV. COMPRESSED COMMUNICATION ALGORITHM: BOUNDED RELATIVE COMPRESSION ERROR

In this section, we use the compressors with bounded relative compression error to design a communication-efficient distributed algorithm and analyze its convergence properties.

A. Algorithm Description

To solve (1), various distributed algorithms have been proposed. For example, Yi et al. [9] proposed the following distributed primal–dual algorithm:

$$x_{i,k+1} = x_{i,k} - \eta \left(\alpha \sum_{j=1}^{n} L_{ij} x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k}) \right) \quad (8a)$$

$$v_{i,k+1} = v_{i,k} + \eta\beta \sum_{j=1}^{n} L_{ij} x_{j,k} \quad (8b)$$

where α, β, and η are positive algorithm parameters, and $x_{i,k} \in \mathbb{R}^d$ is agent i's estimation of the solution to the problem (1) at the kth iteration.

To implement the algorithm (8), at each iteration, each agent j needs to exactly communicate the vector-valued variable $x_{j,k}$ with its neighbors, which requires significant amount of data exchange especially when the dimension d is large. However, in practice, communication bandwidth and power are limited, which motivates the author of this article to consider communication-efficient distributed algorithms. We use communication compression to improve communication efficiency. As mentioned in Section I, directly combining the algorithm (8) and communication compression, i.e., using the compressed variable $\mathcal{C}(x_{j,k})$ to replace $x_{j,k}$ in (8), does not work due to the compression error. To reduce the compression error, an auxiliary variable $a_{j,k} \in \mathbb{R}^d$ is introduced. The difference $x_{j,k} - a_{j,k}$ instead of $x_{j,k}$ is compressed and communicated, and then is added back to $a_{j,k}$ for replacing $x_{j,k}$ in (8). Then, we have the following algorithm:

$$x_{i,k+1} = x_{i,k} - \eta \left(\alpha \sum_{j=1}^{n} L_{ij} \hat{x}_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k}) \right) \quad (9a)$$

$$v_{i,k+1} = v_{i,k} + \eta\beta \sum_{j=1}^{n} L_{ij} \hat{x}_{j,k} \quad (9b)$$

where

$$\hat{x}_{i,k} = a_{i,k} + \mathcal{C}(x_{i,k} - a_{i,k}). \quad (10)$$

Although in the algorithm (9), the compressor error can be reduced, at each iteration each agent j still needs to exactly communicate the vector-valued variable $a_{j,k}$ due to the summation term $\sum_{j=1}^{n} L_{ij} a_{j,k}$ inside (9). Thus, the algorithm (9) does not enjoy the benefits of compression. To overcome that, another auxiliary variable $b_{j,k} \in \mathbb{R}^d$ is introduced to calculate $\sum_{j=1}^{n} L_{ij} a_{j,k}$. The proposed algorithm is presented in pseudocode as Algorithm 1, which is a communication-efficient algorithm since each agent j only communicates the compressed variable $q_{j,k}$ with its neighbors. Noting that $a_{i,0} = b_{i,0} = \mathbf{0}_d$, by mathematical induction, it is straightforward to check that $b_{i,k} = a_{i,k} - \sum_{j=1}^{n} L_{ij} a_{j,k}, \quad \forall i \in [n]$. Then, (11c) and (11d), respectively, can be rewritten as (9a) and (9b). The same idea to use auxiliary variables to reduce the compression error and to implement communication compression has also been used in the literature, e.g., [18], [31].

To end this section, we would like to briefly explain why the compression error is reduced in Algorithm 1 when the class of compressors satisfying Assumption 1 is used. From (4) and (10), we have

$$\mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \hat{x}_{i,k}\|^2] = \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - a_{i,k} - \mathcal{C}(x_{i,k} - a_{i,k})\|^2]$$

$$\leq r_0 \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - a_{i,k}\|^2]. \quad (12)$$

From the proof of Theorem 1 given in Appendix B, we know that $\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - a_{i,k}\|^2]$ converges to zero. Therefore, the compression error is reduced.

B. Convergence Analysis

In this section, we provide convergence analysis for both scenarios without and with Assumption 7. We first have the following convergence result.

Algorithm 1: A Communication-Efficient Distributed Algorithm Using Compressors With Bounded Relative Compression Error.

1: **Input**: positive parameters α, β, η, and ψ.
2: **Initialize**: $x_{i,0} \in \mathbb{R}^d$, $a_{i,0} = b_{i,0} = v_{i,0} = \mathbf{0}_d$, and $q_{i,0} = \mathcal{C}(x_{i,0})$, $\forall i \in [n]$.
3: **for** $k = 0, 1, \dots$ **do**
4:      **for** $i = 1, \dots, n$ in parallel **do**
5:          Broadcast $q_{i,k}$ to $\mathcal{N}_i$ and receive $q_{j,k}$ from $j \in \mathcal{N}_i$.
6:          Update

$$a_{i,k+1} = a_{i,k} + \psi q_{i,k} \tag{11a}$$

$$b_{i,k+1} = b_{i,k} + \psi \left(q_{i,k} - \sum_{j=1}^{n} L_{ij} q_{j,k} \right) \tag{11b}$$

$$x_{i,k+1} = x_{i,k} - \eta\alpha \left(a_{i,k} - b_{i,k} + \sum_{j=1}^{n} L_{ij} q_{j,k} \right)$$
$$- \eta(\beta v_{i,k} + \nabla f_i(x_{i,k})) \tag{11c}$$

$$v_{i,k+1} = v_{i,k} + \eta\beta \left(a_{i,k} - b_{i,k} + \sum_{j=1}^{n} L_{ij} q_{j,k} \right) \tag{11d}$$

$$q_{i,k+1} = \mathcal{C}(x_{i,k+1} - a_{i,k+1}). \tag{11e}$$

7:      **end for**
8: **end for**
9: **Output**: $\{x_{i,k}\}$.

Theorem 1: Suppose that Assumptions 1 and 4–6 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 1 with $\alpha = \kappa_1 \beta$, $\beta > \kappa_2$, $\eta \in (0, \kappa_3)$, and $\psi \in (0, 1/r]$, where κ_1, κ_2, and κ_3 are positive constants given in Appendix B. Then, for any $T \in \mathbb{N}_0$

$$\sum_{k=0}^{T} \sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(1) \tag{13a}$$

$$\mathbf{E}_{\mathcal{C}}[f(\bar{x}_T) - f^*] = \mathcal{O}(1) \tag{13b}$$

where $\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{i,k}$.

Proof: We use Lyapunov analysis to prove this theorem. More specifically, we first appropriately design a function V_k which contains terms $\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2]$, $\mathbf{E}_{\mathcal{C}}[f(\bar{x}_k) - f^*]$, and $\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - a_{i,k}\|^2]$ describing consensus, optimization, and compression errors, respectively. We then prove that $\mathbf{E}_{\mathcal{C}}[V_k]$ is nonincreasing by showing that the difference $\mathbf{E}_{\mathcal{C}}[V_k - V_{k+1}]$ can be lower bounded by $\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + \|x_{i,k} - a_{i,k}\|^2 + \|\nabla f(\bar{x}_k)\|^2]$. We finally show that V_k is nonnegative and get (13a)–(13b) by summarizing the inequalities containing the difference $\mathbf{E}_{\mathcal{C}}[V_k - V_{k+1}]$. The explicit expressions of the Lyapunov function V_k and the right-hand sides of (13a)–(13b) and the detailed proof are given in Appendix B. ∎

We have several remarks on Theorem 1. First, from (13a), we know that $\min_{k \in [T]}\{\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2]\} = \mathcal{O}(1/T)$. In other words, Algorithm 1 finds a first-order stationary point with the well-known rate $\mathcal{O}(1/T)$, which is the same as that achieved by distributed algorithms with exact communication in the literature, e.g, [5], [6], [7], [8], [9]. Second, from (13b), we know that the cost difference

between the global optimum and the resulting stationary point is bounded. Third, it should be pointed out that the settings on the parameters α, β, and η are just sufficient conditions. With some modifications of the proofs, other forms of settings for these parameters still can guarantee the same type of convergence result. Fourth, observe that the definitions of κ_1 and κ_2 given in Appendix B are independent of the parameters related to the compressors. Therefore, the choice of the parameters α and β is independent of the compressors. Finally, the proof of Theorem 1 is inspired by the proof of Theorem 1 in [9]. However, due to the compressed communication, a different Lyapunov function is appropriately designed and the details are also different.

Then, with Assumption 7, the following result states that Algorithm 1 can linearly find a global optimum.

Theorem 2: Suppose that Assumptions 1 and 4–7 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 1 with the same α, β, η, and ψ given in Theorem 1. Then, for any $k \in \mathbb{N}_0$

$$\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*] = \mathcal{O}((1 - \epsilon)^k) \tag{14}$$

where ϵ is a constant in (0,1) given in Appendix C. ∎

Proof: This proof is based on the proof of Theorem 1. From the P–Ł condition, we know that $\mathbf{E}_{\mathcal{C}}[\|\nabla f(\bar{x}_k)\|^2]$ can be lower bounded by $\mathbf{E}_{\mathcal{C}}[f(\bar{x}_k) - f^*]$, which further implies that the difference $\mathbf{E}_{\mathcal{C}}[V_k - V_{k+1}]$ can be lower bounded by $\mathbf{E}_{\mathcal{C}}[V_k]$. Therefore, $\mathbf{E}_{\mathcal{C}}[V_k]$ exponentially decreases to zero. Thus, (14) holds. The explicit expression of the right-hand side of (14) and the detailed proof are given in Appendix C. ∎

We have several remarks on Theorem 2. First, observe that Algorithm 1 uses the same parameters for the cases without and with the P–Ł condition in Theorems 1 and 2, respectively. As a result, it is not needed to check the P–Ł condition before implementing Algorithm 1, which is important since it is normally difficult to check that condition. Second, compared to [31] which used the same type of compressors and established linear convergence under the condition that the global cost function is strongly convex, we show linear convergence under the weaker P–Ł condition and only use a half number of compression and communication operations per iteration since in the algorithm proposed in [31] each agent needs to communicate two compressed variables with its neighbors. Third, compared to Liu et al. [18], [19] which used unbiased compressors with bounded relative compression error and established linear convergence under the condition that each local cost function is strongly convex, we use the more general compressors and the weaker P–Ł condition to show linear convergence. Finally, compared to [17], which used unbiased compressors with bounded relative compression error but only achieved sublinear convergence under the condition that each local cost function is strongly convex, we not only use the more general compressors and the weaker P–Ł condition but also show strictly faster convergence.

V. ERROR FEEDBACK BASED COMPRESSED COMMUNICATION ALGORITHM: BOUNDED RELATIVE COMPRESSION ERROR

In this section, we extend Algorithm 1 to error feedback version for biased compressors particularly.

A. Algorithm Description

The error feedback based communication-efficient distributed algorithm is presented in pseudocode as Algorithm 2. Without

Algorithm 2: An Error Feedback Based Compressed Communication Algorithm.

1: **Input**: positive parameters α, β, η, ψ, and σ.
2: **Initialize**: $x_{i,0} \in \mathbb{R}^d$, $a_{i,0} = b_{i,0} = e_{i,0} = v_{i,0} = \mathbf{0}_d$, and $q_{i,0} = \hat{q}_{i,0} = \mathcal{C}(x_{i,0})$, $\forall i \in [n]$.
3: **for** $k = 0, 1, \dots$ **do**
4:     **for** $i = 1, \dots, n$ in parallel **do**
5:        Broadcast $q_{i,k}$ and $\hat{q}_{i,k}$ to $\mathcal{N}_i$ and receive $q_{j,k}$ and $\hat{q}_{j,k}$ from $j \in \mathcal{N}_i$.
6:        Update

$$a_{i,k+1} = a_{i,k} + \psi q_{i,k} \tag{15a}$$

$$b_{i,k+1} = b_{i,k} + \psi \left(q_{i,k} - \sum_{j=1}^{n} L_{ij} q_{j,k} \right) \tag{15b}$$

$$x_{i,k+1} = x_{i,k} - \eta\alpha \left(a_{i,k} - b_{i,k} + \sum_{j=1}^{n} L_{ij} \hat{q}_{j,k} \right)$$
$$- \eta(\beta v_{i,k} + \nabla f_i(x_{i,k})) \tag{15c}$$

$$v_{i,k+1} = v_{i,k} + \eta\beta \left(a_{i,k} - b_{i,k} + \sum_{j=1}^{n} L_{ij} \hat{q}_{j,k} \right) \tag{15d}$$

$$q_{i,k+1} = \mathcal{C}(x_{i,k+1} - a_{i,k+1}) \tag{15e}$$

$$e_{i,k+1} = \sigma e_{i,k} + x_{i,k} - a_{i,k} - \hat{q}_{i,k} \tag{15f}$$

$$\hat{q}_{i,k+1} = \mathcal{C}(\sigma e_{i,k+1} + x_{i,k+1} - a_{i,k+1}). \tag{15g}$$

7:     **end for**
8: **end for**
9: **Output**: $\{x_{i,k}\}$.

ambiguity, we denote

$$\hat{x}_{i,k} = a_{i,k} + \hat{q}_{i,k} \tag{16}$$

then (15c) and (15d), respectively, can be written as (9a) and (9b) since $b_{i,k} = a_{i,k} - \sum_{j=1}^{n} L_{ij} a_{j,k}$, $\forall i \in [n]$. Therefore, Algorithm 2 also is a communication-efficient extension of the distributed primal–dual algorithm (8).

Compared to Algorithm 1, Algorithm 2 has two new variables $\hat{q}_{i,k}$ and $e_{i,k}$, which are used to estimate the biased compression error and to accumulate the biased compression errors, respectively. Then each agent can use $\hat{q}_{i,k}$ to correct the bias induced by the biased compressors.[3] However, compared to Algorithm 1, there are twice number of compression and communication operations per iteration in Algorithm 2.

B. Convergence Analysis

Similar to Theorem 1, we first have the following sublinear convergence result for Algorithm 2 without Assumption 7.

Theorem 3: Suppose that Assumptions 1 and 4–6 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 2 with $\alpha = \kappa_1\beta$, $\beta > \kappa_2$, $\eta \in (0, \check{\kappa}_3)$, $\sigma \in (0, \kappa_0)$, and $\psi \in (0, 1/r]$, where

---

[3]For unbiased compressors, it is unnecessary to consider error feedback since $\mathbf{E}_{\mathcal{C}}[e_{i,k}] = \mathbf{0}_d$.

κ_1, κ_2 and κ_0, $\check{\kappa}_3$ are positive constants given in Appendices B and D, respectively. Then, for any $T \in \mathbb{N}_0$

$$\sum_{k=0}^{T} \sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(1) \tag{17a}$$

$$\mathbf{E}_{\mathcal{C}}[f(\bar{x}_T) - f^*] = \mathcal{O}(1). \tag{17b}$$

Proof: This proof is similar to the proof of Theorem 1 but uses a different Lyapunov function W_k. Due to space limitations, the explicit expressions of the Lyapunov function W_k and the right-hand sides of (17a)–(17b) and the detailed proof are given in [45]. ∎

Similar to Theorem 2, we then have the following linear convergence result for Algorithm 2 with Assumption 7.

Theorem 4: Suppose that Assumptions 1 and 4–7 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 2 with the same α, β, η, σ, and ψ given in Theorem 3. Then, for any $k \in \mathbb{N}_0$

$$\sum_{i=1}^{n} \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*] = \mathcal{O}((1 - \check{\epsilon})^k) \tag{18}$$

where $\check{\epsilon}$ is a constant in $(0, 1)$ given in Appendix E. ∎

Algorithm 3: A Communication-Efficient Distributed Algorithm Using Compressors With Bounded Absolute Compression Error

1: **Input**: positive parameters α, β, η, and a positive scaling sequence $\{s_k\}$.
2: **Initialize**: $x_{i,0} \in \mathbb{R}^d$, $\hat{x}_{i,-1} = y_{i,-1} = v_{i,0} = \mathbf{0}_d$, and $q_{i,0} = \mathcal{C}(x_{i,0}/s_0)$, $\forall i \in [n]$.
3: **for** $k = 0, 1, \dots$ **do**
4:     **for** $i = 1, \dots, n$ in parallel **do**
5:        Broadcast $q_{i,k}$ to $\mathcal{N}_i$ and receive $q_{j,k}$ from $j \in \mathcal{N}_i$.
6:        Update

$$\hat{x}_{i,k} = \hat{x}_{i,k-1} + s_k q_{i,k} \tag{19a}$$

$$y_{i,k} = y_{i,k-1} + s_k q_{i,k} - s_k \sum_{j=1}^{n} L_{ij} q_{j,k} \tag{19b}$$

$$x_{i,k+1} = x_{i,k} - \eta\alpha(\hat{x}_{i,k} - y_{i,k})$$
$$- \eta(\beta v_{i,k} + \nabla f_i(x_{i,k})) \tag{19c}$$

$$v_{i,k+1} = v_{i,k} + \eta\beta(\hat{x}_{i,k} - y_{i,k}) \tag{19d}$$

$$q_{i,k+1} = \mathcal{C}((x_{i,k+1} - \hat{x}_{i,k})/s_{k+1}). \tag{19e}$$

7:     **end for**
8: **end for**
9: **Output**: $\{x_{i,k}\}$.

Proof: This proof is similar to the proof of Theorem 2 but uses the Lyapunov function W_k as used in the proof of Theorem 3. Due to space limitations, the expression of the right-hand side of (18) and the detailed proof are given in [45].

VI. COMPRESSED COMMUNICATION ALGORITHM: BOUNDED ABSOLUTE COMPRESSION ERROR

In this section, we use the compressors with bounded absolute compression error to design a communication-efficient distributed algorithm and analyze its convergence properties in various setups.

A. Algorithm Description

The communication-efficient distributed algorithm is presented in pseudocode as Algorithm 3.

By mathematical induction, it is straightforward to check that $y_{i,k} = \hat{x}_{i,k} - \sum_{j=1}^{n} L_{ij}\hat{x}_{j,k}, \; \forall i \in [n]$. Therefore, (19c) and (19d) can be rewritten as (9a) and (9b), respectively. Therefore, Algorithm 3 also is a communication-efficient extension of the distributed primal–dual algorithm (8). Moreover, same as Algorithm 1, in Algorithm 3 each agent only communicates one compressed variable with its neighbors. The difference between Algorithms 1 and 3 is that they use different types of compressors.

B. Convergence Analysis

We first analyze the performance of Algorithm 3 when the class of compressors satisfying Assumption 2 is used. Before stating the convergence results, we would like to briefly explain why this algorithm works. From (5), (19a), and (19e), we have

$$\mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \hat{x}_{i,k}\|_p^2]$$

$$= \mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \hat{x}_{i,k-1} - s_k\mathcal{C}((x_{i,k} - \hat{x}_{i,k-1})/s_k)\|_p^2]$$

$$= \mathbf{E}_{\mathcal{C}}[s_k^2\|(x_{i,k} - \hat{x}_{i,k-1})/s_k - \mathcal{C}((x_{i,k} - \hat{x}_{i,k-1})/s_k)\|_p^2]$$

$$\leq Cs_k^2. \tag{20}$$

If we let s_k exponentially decrease to zero, then the error caused by the compressed communication is neglectable. In this case, Algorithm 3 using the second class of compressors can have comparable convergence properties as the corresponding algorithm with exact communication, i.e., (8).

Similar to Theorem 1, we have the following sublinear convergence result.

Theorem 5: Suppose that Assumptions 2 and 4–6 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 3 with $\alpha = \kappa_1\beta$, $\beta > \kappa_2, \eta \in (0, \tilde{\kappa}_3)$, and $s_k = s_0\gamma^k$, where κ_1 and κ_2 are positive constants given in Appendix B, $\tilde{\kappa}_3$ a positive constant given in Appendix F, s_0 is an arbitrary positive constant, and γ is an arbitrary constant in (0,1). Then, for any $T \in \mathbb{N}_0$

$$\sum_{k=0}^{T}\sum_{i=1}^{n}\mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2] = \mathcal{O}(1) \tag{21a}$$

$$\mathbf{E}_{\mathcal{C}}[f(\bar{x}_T) - f^*] = \mathcal{O}(1). \tag{21b}$$

Proof: This proof is similar to the proof of Theorem 1 but uses the nonnegative function U_k which contains terms describing consensus and optimization errors and is given in Appendix B. We show that the difference $\mathbf{E}_{\mathcal{C}}[U_k - U_{k+1}]$ can be lower bounded by $\sum_{i=1}^{n}\mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 - \|x_{i,k} - \hat{x}_{i,k}\|^2 + \|\nabla f(\bar{x}_k)\|^2]$. From (20) and $s_k = s_0\gamma^k$, we can get (21a)–(21b) by summarizing the inequalities containing the difference $\mathbf{E}_{\mathcal{C}}[U_k - U_{k+1}]$. Due to space limitations, the explicit expressions of the right-hand sides of (21a)–(21b) and the detailed proof are given in [45]. ∎

Similar remarks as those after Theorem 1 are valid for Theorem 5. Moreover, we would like to point out that the choice of the parameter γ is also independent of the compressors since the definition of $\tilde{\kappa}_3$ given in Appendix F is independent of the parameters related to the compressors.

Similar to Theorem 2, we then have the following linear convergence result for Algorithm 3 when the class of compressors satisfying Assumption 2 is used.

Theorem 6: Suppose that Assumptions 2 and 4–7 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 3 with the same α, β, η, and s_k given in Theorem 5. Then, for any $k \in \mathbb{N}_0$

$$\sum_{i=1}^{n}\mathbf{E}_{\mathcal{C}}[\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*] = \mathcal{O}((1 - \tilde{\epsilon})^k) \tag{22}$$

where $\tilde{\epsilon}$ is a constant in (0,1) given in Appendix G.

Proof: This proof is based on the proof of Theorem 5. From the P–Ł condition, we know that $\mathbf{E}_{\mathcal{C}}[\|\nabla f(\bar{x}_k)\|^2]$ can be lower bounded by $\mathbf{E}_{\mathcal{C}}[f(\bar{x}_k) - f^*]$, which further implies that the difference $\mathbf{E}_{\mathcal{C}}[U_k - U_{k+1}]$ can be lower bounded by $\mathbf{E}_{\mathcal{C}}[U_k - \sum_{i=1}^{n}\|x_{i,k} - \hat{x}_{i,k}\|^2]$. Then, combining this, (20), and $s_k = s_0\gamma^k$, we can get that $\mathbf{E}_{\mathcal{C}}[U_k]$ exponentially decreases to zero. Thus, (22) holds. Due to space limitations, the explicit expression of the right-hand side of (22) and the detailed proof are given in [45]. ∎

Compared to [32] which used the same class of compressors satisfying Assumption 2, Theorem 6 shows that a global optimum can be precisely found with a linear convergence rate under the P–Ł condition. In contrast, although Khirirat et al. [32] assumed the stronger strong convexity assumption and also showed that convergence rate is linear, the parallel algorithms proposed in [32] only converged to a neighbor of the unique optimal point.

We also have the following linear convergence result for Algorithm 3 when the class of compressors satisfying Assumption 3 is used.

Theorem 7: Suppose that Assumptions 3–7 hold and a lower bound on the P–Ł constant ν is known in advance. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 3 with $\alpha = \kappa_1\beta$, $\beta > \kappa_2$, $\eta \in (0, \hat{\kappa}_3)$, and $s_k = s_0\gamma^k$, where κ_1 and κ_2 are positive constants given in Appendix B, $\hat{\kappa}_3, s_0$, and γ are positive constants given in Appendix H with $\gamma \in (0, 1)$. Then, for any $k \in \mathbb{N}_0$

$$\sum_{i=1}^{n}(\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^*) = \mathcal{O}(\gamma^k). \tag{23}$$

Proof: This proof also uses the nonnegative function U_k as used in the proof of Theorem 5. Note that the inequality (6) in Assumption 3 only holds locally. We use mathematical induction to prove that $\max_{i \in [n]}\|(x_{i,k} - \hat{x}_{i,k-1})/s_k\|_p^2 \leq 1$ and U_k/s_k^2 is globally bounded. Thus, (23) holds. Due to space limitations, the explicit expression of the right-hand side of (23) and the detailed proof are given in [45]. ∎

We have several remarks on Theorem 7. First, compared to Theorems 2, 4, and 6, Theorem 7 needs a lower bound on the P–Ł constant ν to be known in advance, which is used to design the parameters s_0 and γ, as shown in Appendix A8. This is a potential drawback since this constant is normally unknown due to the difficulty to check the P–Ł condition. However, for strongly convex cost functions, this is not a drawback since if a function is strongly convex with convex parameter ν, then it also satisfies the P–Ł condition with the same constant ν. Second, linear convergence has also been established in [33], which used the same type of compressors. However, Zhang et al. [33] assumed that each local cost function is strongly convex, which is stronger than the condition that the global cost function satisfies the P–Ł condition as used in Theorem 7, and required that the absolute compression error satisfies an inequality determined by the number of agents and the communication network, which is not needed in Theorem 7. Moreover, Zhang et al. [33] required an unpractical condition that the unique optimal point needs to be known a priori to design algorithm parameters, which is a potential drawback. Third, compared to [26], [27], [28], and [29] which used the standard uniform quantizer with dynamic

quantization level and established linear convergence under the condition that each local cost function is strongly convex, we use the more general compressors and the weaker P–Ł condition to show linear convergence. Finally, compared to [30] which used the standard uniform quantizer with fixed quantization level and established linear convergence under the assumption that each local cost function is quadratic and the global cost function is strongly convex, we not only use the more general compressors but also consider the more general nonconvex functions satisfying the weaker P–Ł condition.

To end this section, we would like to clarify that although this article considers three different general classes of compressors, it is not this article's goal to study which specific compressor or general class of compressors has better performance. Moreover, although this article proposes three communication-efficient distributed algorithms, it is not this article's goal either to investigate which algorithm has better performance.

VII. SIMULATIONS

In this section, we verify and illustrate the theoretical results through numerical simulations. We consider the nonconvex distributed binary classification problem as studied in [8], [9], and [11], which is formulated as the optimization problem (1) with each component function f_i being given by

$$f_i(x) = \frac{n}{m} \sum_{l=1}^{m_i} \left((1 - y_{il}) \log \left(1 + e^{x^\top z_{il}} \right) \right.$$

$$\left. + y_{il} \log \left(1 + e^{-x^\top z_{il}} \right) \right) + \sum_{s=1}^{d} \frac{\lambda \mu [x]_s^2}{1 + \mu [x]_s^2}$$

where $m = \sum_{i=1}^{n} m_i$, m_i is the number of observations held privately by agent i, $z_{il} \in \mathbb{R}^p$ is the lth observation with label $y_{il} \in \{0, 1\}$ owned by agent i, λ and μ are regularization parameters, and $[x]_s$ is the sth coordinate of $x \in \mathbb{R}^d$. All settings for cost functions and the communication graph are the same as those described in [8] and [9]. Specifically, $n = 20$, $d = 50$, $m_i = 200$, $\lambda = 0.001$, and $\mu = 1$. The graph used in the simulation is the random geometric graph and the graph parameter is set to be 0.5. We independently and randomly generate m data points.

We consider the following five compressors.
1) Unbiased l-bits quantizer [18]

$$\mathcal{C}_1(x) = \frac{\|x\|_\infty}{2^{l-1}} \text{sign}(x) \circ \left\lfloor \frac{2^{l-1} |x|}{\|x\|_\infty} + \varpi \right\rfloor$$

where $\text{sign}(\cdot)$, $|\cdot|$, and $\lfloor \cdot \rfloor$ are the elementwise sign, absolute, and floor functions, respectively, $\circ$ denotes the Hadamard product, and ϖ is a random perturbation vector uniformly sampled from $[0, 1]^d$. This compressor is unbiased and satisfies Assumption 1 with $r = 1 + r_1$, $\varphi = 1/(1 + r_1)$, and $r_1 = d/4^l$. As pointed out in [33], transmitting $\mathcal{C}_1(x)$ needs $(l + 1)d + b_1$ bits if a scalar can be transmitted with b_1 bits with sufficient precision since only $\|x\|_\infty$, $\text{sign}(x)$, and the positive integer in the bracket need to be transmitted. In this section, we choose $l = 2$ and $b_1 = 64$.
2) Greedy (Top-k) sparsifier [13]

$$\mathcal{C}_2(x) = \sum_{s=1}^{k} [x]_{i_s} \mathbf{e}_{i_s}$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ is the standard basis of $\mathbb{R}^d$ and $i_1, \ldots, i_k$ are the indices of largest k coordinates in magnitude of x. This compressor is biased but contractive. Therefore, it satisfies Assumption 1 with $r = 1$ and $\varphi = k/d$. Moreover, it also satisfies Assumption 3 with $p \geq 2$ and $\varphi = k/d$. Transmitting $\mathcal{C}_2(x)$ needs kb_1 bits since only k scalars need to be transmitted. In this section, we choose $k = 10$.
3) Norm-sign compressor

$$\mathcal{C}_3(x) = \frac{\|x\|_\infty}{2} \text{sign}(x).$$

This compressor is biased and noncontractive but satisfies Assumption 1 with $r = d/2$ and $\varphi = 1/d^2$, see [31]. It also satisfies Assumption 3 with $p = \infty$ and $\varphi = 0.5$. Transmitting $\mathcal{C}_3(x)$ needs $2d + b_1$ bits since only $\|x\|_\infty$ and $\text{sign}(x)$ need to be transmitted.
4) Standard uniform quantizer

$$\mathcal{C}_4(x) = \Delta \left\lfloor \frac{x}{\Delta} + \frac{\mathbf{1}_d}{2} \right\rfloor$$

where Δ is a positive integer. This compressor satisfies Assumption 2 with $p = \infty$ and $C = \Delta^2/4$. Moreover, it also satisfies Assumption 3 with $p = \infty$ and $\varphi = 0.5$ when $\Delta = 1$. Transmitting $\mathcal{C}_4(x)$ needs db_2 bits if b_2 bits are allocated to transmit an integer. In this section, we choose $\Delta = 1$ and $b_2 = 8$.
5) 1-bit binary quantizer [33]

$$\mathcal{C}_5(x) = \text{col}(Q_1([x]_1), \ldots, Q_1([x]_d))$$

where $Q_1([x]_s) = 0.5$ for $[x]_s \geq 0$ and $Q_1([x]_s) = -0.5$ otherwise. This compressor satisfies Assumption 3 with $p = \infty$ and $\varphi = 0.5$. Transmitting $\mathcal{C}_5(x)$ needs d bits since for each coordinate only two symbols need to be transmitted.
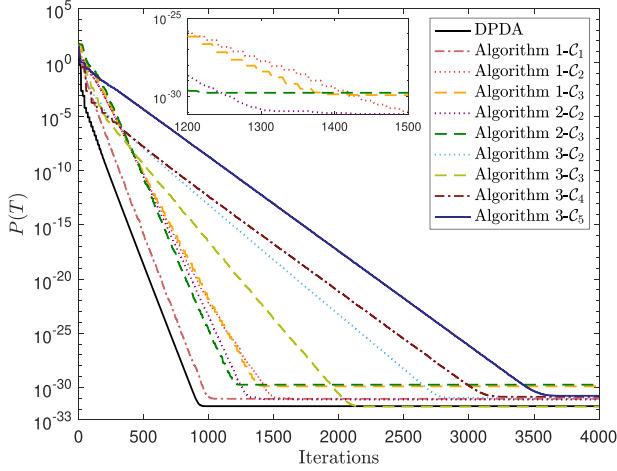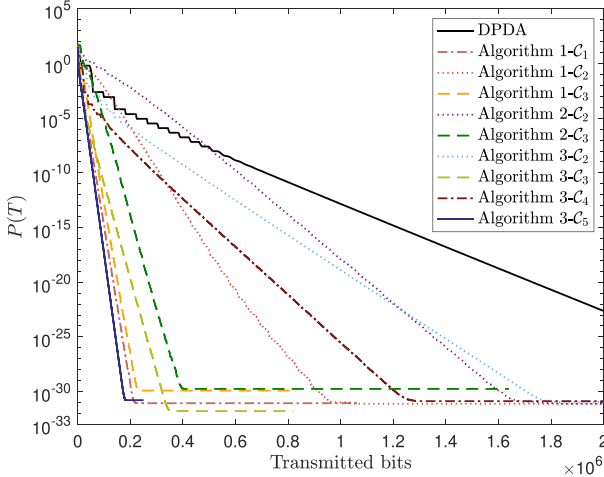
We implement Algorithm 1 using $\mathcal{C}_1$–$\mathcal{C}_3$, Algorithm 2 using $\mathcal{C}_2$ and $\mathcal{C}_3$, and Algorithm 3 using $\mathcal{C}_2$–$\mathcal{C}_5$. Note that, to the best of our knowledge, in the literature there are no other similar communication-efficient distributed algorithms for distributed nonconvex optimization as ours. Therefore, we only compare the proposed communication-efficient distributed algorithms with their uncompressed counterpart, i.e., the distributed primal–dual algorithm (8), which is denoted as DPDA. It is straightforward to see that each agent sends db_1 bits per iteration when implementing DPDA. All the hyperparameters used in the experiment are tuned manually and given in [45, Table I] due to space limitations.

We use

$$P(T) = \min_{k \in [T]} \left\{ \|\nabla f(\bar{x}_k)\|^2 + \frac{1}{n} \sum_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 \right\}$$
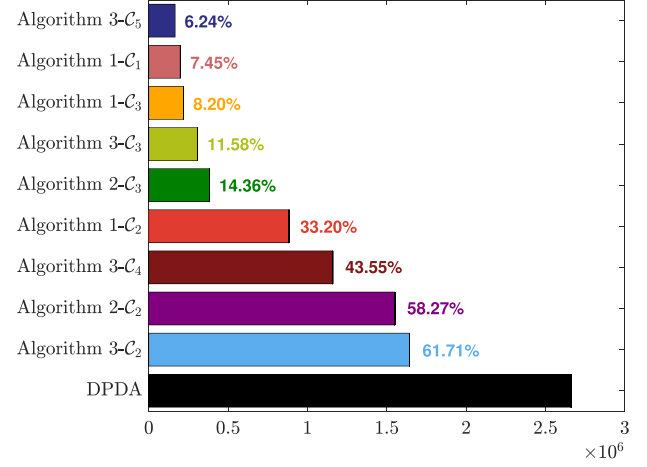
to measure the performance of each algorithm. We plot the convergence of $P(T)$ with respect to both number of iterations and bits transmitted between two neighbor agents for the above algorithm and compressor combinations with the same initial condition, as shown in Figs. 1 and 2, respectively. Moreover, the comparison of transmitted bits for different algorithm and compressor combinations to reach $P(T) \leq 10^{-30}$ is provided in Fig. 3. We highlight the following observations.
1) From Fig. 1, we can see that all of the algorithm and compressor combinations have comparable convergence

Fig. 1. Evolutions of $P(T)$ with respect to the number of iterations.

Fig. 2. Evolutions of $P(T)$ with respect to the number of transmitted bits.

Fig. 3. Transmitted bits for different algorithm and compressor combinations to reach $P(T) \leq 10^{-30}$.

Algorithm 1-$\mathcal{C}_1$, Algorithm 1-$\mathcal{C}_3$, Algorithm 3-$\mathcal{C}_3$, and Algorithm 2-$\mathcal{C}_3$ converge significantly faster than DPDA. For example, it is illustrated in Fig. 3 that Algorithm 3-$\mathcal{C}_5$ only needs 6.24% of the bits used by DPDA to reach a specific level of error.

5) From Fig. 2, we can also see that Algorithm 1-$\mathcal{C}_2$ (Algorithm 1-$\mathcal{C}_3$) converges faster than its error feedback version, i.e., Algorithm 2-$\mathcal{C}_2$ (Algorithm 2-$\mathcal{C}_3$), when comparing their performances based on the number of transmitted bits, which reveals the potential drawback of using error feedback to correct the bias induced by the biased compressors.

VIII. CONCLUSION

In this article, we studied communication compression for distributed nonconvex optimization. We used three general classes of compressors to design three communication-efficient distributed primal–dual algorithms. We showed that the proposed algorithms can achieve comparable convergence results as state-of-the-art algorithms, although the communication is compressed. Interesting directions for future work include considering more general network topologies, reducing communication complexity through periodic communication, exploring rules for choosing an appropriate compressor for high efficiency, and studying how the important parameters, such as compressor parameters, Lipschitz constant, and network connectivity, affect the convergence rate.

APPENDIX

A. Useful Lemmas

The following results are used in the proofs.

Lemma 1 [46, eq. (5.4.21), p. 333]: For any $x \in \mathbb{R}^d$, it holds that $\|x\|_p \leq \hat{d}\|x\|$ and $\|x\| \leq \tilde{d}\|x\|_p$, where $\hat{d} = d^{\frac{1}{p} - \frac{1}{2}}$ and $\tilde{d} = 1$ when $p \in [1, 2]$, and $\hat{d} = 1$ and $\tilde{d} = d^{\frac{1}{2} - \frac{1}{p}}$ when $p > 2$.

Lemma 2: For any $x, y \in \mathbb{R}^d$ and $a, b > 0$ satisfying $ab = \frac{1}{4}$, it holds that

$$x^\top y \leq a\|x\|^2 + b\|y\|^2. \tag{24}$$

This lemma is a direct extension of the Cauchy–Schwarz inequality.

speeds as the corresponding algorithm with exact communication, i.e., DPDA, which is consistent with the theoretical results. Especially, Algorithm 1-$\mathcal{C}_1$ and DPDA have almost the same convergence speed.

2) From Fig. 1, we can also see that Algorithm 1-$\mathcal{C}_3$ (Algorithm 2-$\mathcal{C}_3$) has almost the same convergence speed as Algorithm 1-$\mathcal{C}_2$ (Algorithm 2-$\mathcal{C}_2$), and Algorithm 3-$\mathcal{C}_3$ has faster convergence speed than Algorithm 3-$\mathcal{C}_2$. Therefore, noncontractive compressors, e.g., $\mathcal{C}_3$, can converge faster than contractive compressors, e.g., $\mathcal{C}_2$.

3) From the zoomed figure in Fig. 1, we can see that the error feedback method Algorithm 2-$\mathcal{C}_2$ (Algorithm 2-$\mathcal{C}_3$) has faster convergence speed than Algorithm 1-$\mathcal{C}_2$ (Algorithm 1-$\mathcal{C}_3$), which demonstrates the benefit of using error feedback to correct the bias induced by the biased compressors.

4) From Fig. 2, we can see that the communication-efficient algorithms converge faster than their exact-communication counterpart when comparing their performances based on the number of bits that each agents communicates, which shows the effectiveness of the proposed algorithms. Especially, Algorithm 3-$\mathcal{C}_5$,

Lemma 3: Let L be the Laplacian matrix of an undirected and connected graph $\mathcal{G}$ with n agents and $K_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Then L and K_n are positive semidefinite, $\text{null}(L) = \text{null}(K_n) = \{\mathbf{1}_n\}$, $L \leq \rho(L)\mathbf{I}_n$, $\rho(K_n) = 1$

$$K_n L = L K_n = L \tag{25a}$$

$$0 \leq \rho_2(L)K_n \leq L \leq \rho(L)K_n. \tag{25b}$$

Moreover, there exists an orthogonal matrix $[r\ R] \in \mathbb{R}^{n \times n}$ with $r = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $R \in \mathbb{R}^{n \times (n-1)}$ such that

$$L = \begin{bmatrix} r & R \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \begin{bmatrix} r^\top \\ R^\top \end{bmatrix} \tag{26a}$$

$$PL = LP = K_n \tag{26b}$$

$$\rho^{-1}(L)\mathbf{I}_n \leq P \leq \rho_2^{-1}(L)\mathbf{I}_n \tag{26c}$$

where $\Lambda_1 = \text{diag}([\lambda_2, \ldots, \lambda_n])$ with $0 < \lambda_2 \leq \cdots \leq \lambda_n$ being the nonzero eigenvalues of the Laplacian matrix L, and

$$P = \begin{bmatrix} r & R \end{bmatrix} \begin{bmatrix} \lambda_n^{-1} & 0 \\ 0 & \Lambda_1^{-1} \end{bmatrix} \begin{bmatrix} r^\top \\ R^\top \end{bmatrix}.$$

Proof: From [47, Lemmas 1 and 2], we know that all the results except (26b)–(26c) hold. ∎

From that $[r\ R]$ is an orthogonal matrix, (26a), and the definitions of K_n, P, and Q, it is straightforward to check that (26b)–(26c) hold.

B. Proof of Theorem 1

To prove Theorem 1, we first introduce some constants and notations. Denote the following constants:

$$\kappa_1 \geq \max\left\{\frac{9 + \kappa_4}{2\rho_2(L)}, 1\right\}, \kappa_2 = \max\{\kappa_5, \sqrt{\kappa_6}\}$$

$$\kappa_3 = \min\left\{\frac{\epsilon_1}{\epsilon_2}, \frac{\epsilon_3}{\epsilon_4}, \frac{\epsilon_5}{\epsilon_6}, \frac{\sqrt{\epsilon_8^2 + 4\epsilon_7\epsilon_9} - \epsilon_8}{2\epsilon_9}\right\}$$

$$\kappa_4 > 0, \kappa_5 = \max\left\{\frac{4 + 5L_f^2}{\kappa_4}, \frac{6}{\rho_2(L)}\right\}$$

$$\kappa_6 = \frac{8(\kappa_1 + 1)^2 L_f^2}{\kappa_5\rho_2(L)} + \frac{4L_f^2}{\rho_2^2(L)}$$

$$\epsilon_1 = \frac{\alpha}{2}\rho_2(L) - \frac{1}{4}(9\beta + 4 + 5L_f^2)$$

$$\epsilon_2 = 3L_f^2 + \frac{4(2 + \varphi\psi r)}{\varphi\psi r}(\alpha^2\rho^2(L) + L_f^2)$$
$$+ 2\beta^2 + 1 + 3\alpha^2\rho^2(L)$$

$$\epsilon_3 = \frac{\beta}{2} - 3\rho_2^{-1}(L)$$

$$\epsilon_4 = 2\beta^2\rho(L) + \frac{4(2 + \varphi\psi r)\beta^2\rho^2(L)}{\varphi\psi r} + \rho_2^{-1}(L)$$

$$\epsilon_5 = \frac{1}{8} - \frac{(\alpha + \beta)^2 L_f^2}{\beta^5\rho_2(L)} - \frac{L_f^2}{2\beta^2\rho_2^2(L)}$$

$$\epsilon_6 = \frac{(\alpha + \beta)L_f^2}{2\beta^3\rho_2(L)} + \frac{3L_f^2}{4} + \frac{L_f^2}{2\beta^2\rho_2^2(L)} + \frac{L_f}{2}$$

$$\epsilon_7 = \frac{\varphi\psi r}{2}(1 + \varphi\psi r)$$

$$\epsilon_8 = \frac{1}{2}(\alpha + 2\beta)\rho(L)r_0 + 2\beta r_0$$

$$\epsilon_9 = \frac{(8 + 7\varphi\psi r)\alpha^2\rho^2(L)r_0}{\varphi\psi r} + (2\beta^2 + 1)r_0$$

$$\epsilon_{10} = \frac{\alpha\rho_2(L) - \beta}{2\alpha\rho_2(L)}$$

$$c_1 = \left(\frac{(\alpha + \beta)^2}{\eta\beta^5} + \frac{\alpha + \beta}{2\beta^3}\right)\frac{1}{\rho_2(L)} + \frac{1}{2}$$

$$c_2 = \frac{\eta + 1}{2\eta\beta^2\rho_2^2(L)} + \frac{1}{4}, c_3 = \frac{\varphi\psi r}{2}$$

$$c_4 = c_3 + 2c_3^2 - 4\eta^2(1 + c_3^{-1})\alpha^2\rho^2(L)r_0$$

$$c_5 = \frac{\alpha}{2}\rho_2(L) - \frac{1}{4}(\beta + 4 + 5L_f^2)$$

$$c_6 = 3L_f^2 + 4(1 + c_3^{-1})(\alpha^2\rho^2(L) + L_f^2)$$

$$c_7 = \beta^2 + \frac{1}{2} + \frac{3\alpha^2\rho^2(L)}{2}.$$

Denote $\boldsymbol{x} = \text{col}(x_1, \ldots, x_n)$, $\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$, $\boldsymbol{L} = L \otimes \mathbf{I}_d$, $\boldsymbol{H} = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d)$, $\boldsymbol{K} = K_n \otimes \mathbf{I}_d = \mathbf{I}_{nd} - \boldsymbol{H}$, $\boldsymbol{P} = P \otimes \mathbf{I}_d$, $\bar{\boldsymbol{x}}_k = \mathbf{1}_n \otimes \bar{x}_k = \boldsymbol{H}\boldsymbol{x}_k$, $\boldsymbol{g}_k = \nabla\tilde{f}(\boldsymbol{x}_k)$, $\bar{\boldsymbol{g}}_k = \boldsymbol{H}\boldsymbol{g}_k$, $\boldsymbol{g}_k^0 = \nabla\tilde{f}(\bar{\boldsymbol{x}}_k)$, and $\bar{\boldsymbol{g}}_k^0 = \boldsymbol{H}\boldsymbol{g}_k^0 = \mathbf{1}_n \otimes \nabla f(\bar{x}_k)$. Moreover, without ambiguity, we denote $\mathcal{C}(\boldsymbol{x}) = \text{col}(\mathcal{C}(x_1), \ldots, \mathcal{C}(x_n))$. We also denote

$$U_{1,k} = \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2, U_{2,k} = \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\frac{\alpha+\beta}{\beta}\boldsymbol{P}}^2$$

$$U_{3,k} = \boldsymbol{x}_k^\top \boldsymbol{K}\boldsymbol{P}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right), U_{4,k} = n(f(\bar{x}_k) - f^*)$$

$$U_k = \sum_{i=1}^4 U_{i,k}, V_k = U_k + \|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2$$

$$\hat{U}_k = \|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P}}^2 + n(f(\bar{x}_k) - f^*)$$

$$\hat{V}_k = \hat{U}_k + \|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2$$

$$\boldsymbol{M}_1 = \frac{\alpha}{2}\boldsymbol{L} - \frac{1}{4}(\beta + 4 + 5L_f^2)\boldsymbol{K}$$

$$\boldsymbol{M}_2 = \left(\beta^2 + \frac{1}{2}\right)\boldsymbol{K} + \frac{\beta(\beta - \alpha)}{2}\boldsymbol{L} + \frac{3\alpha^2}{2}\boldsymbol{L}^2.$$

Note that $U_{4,k}$ is well defined since $f^* > -\infty$ as assumed in Assumption 5. To prove Theorem 1, the following lemma is used, which presents a general relation between two consecutive outputs of Algorithm 1.

Lemma 4: Suppose Assumptions 1 and 4–6 hold. Let $\{x_{i,k}\}$ be the sequence generated by Algorithm 1 with $\alpha \geq \beta$ and $\psi \in (0, 1/r]$. Then

$$\mathbf{E}_\mathcal{C}[V_{k+1}] \leq \mathbf{E}_\mathcal{C}\left[V_k - \frac{\eta}{4}\|\bar{\boldsymbol{g}}_k^0\|^2 - \|\boldsymbol{x}_k\|_{\eta(\epsilon_1 - \eta\epsilon_2)\boldsymbol{K}}^2\right.$$
$$\left. - \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\eta(\epsilon_3 - \eta\epsilon_4)\boldsymbol{P}}^2 - \eta(\epsilon_5 - \eta\epsilon_6)\|\bar{\boldsymbol{g}}_k\|^2\right.$$

$$-(\epsilon_7 - \eta\epsilon_8 - \eta^2\epsilon_9)\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2]. \tag{27}$$

Proof: 1) We first introduce some useful equations.

The compact form of (11a), (11b), (9a), (9b), and (11e) is

$$\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \psi\boldsymbol{q}_k \tag{28a}$$

$$\boldsymbol{b}_{k+1} = \boldsymbol{b}_k + \psi(\mathbf{I}_{np} - \boldsymbol{L})\boldsymbol{q}_k \tag{28b}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta(\alpha\boldsymbol{L}\hat{\boldsymbol{x}}_k + \beta\boldsymbol{v}_k + \nabla\tilde{f}(\boldsymbol{x}_k)) \tag{28c}$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta\beta\boldsymbol{L}\hat{\boldsymbol{x}}_k \tag{28d}$$

$$\boldsymbol{q}_{k+1} = \mathcal{C}(\boldsymbol{x}_{k+1} - \boldsymbol{a}_{k+1}). \tag{28e}$$

Denote $\bar{v}_k = \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\boldsymbol{v}_k$. Then, from (28d) and $\sum_{i=1}^n L_{ij} = 0$, we know that $\bar{v}_{k+1} = \bar{v}_k$. This together with the fact that $\sum_{i=1}^n v_{i,0} = \mathbf{0}_d$ implies

$$\bar{v}_k = \mathbf{0}_d. \tag{29}$$

Then, from (28c) and (29), we know that

$$\bar{\boldsymbol{x}}_{k+1} = \bar{\boldsymbol{x}}_k - \eta\bar{\boldsymbol{g}}_k. \tag{30}$$

Noting that $\nabla\tilde{f}$ is Lipschitz-continuous with constant $L_f > 0$ as assumed in Assumption 6, we have

$$\|\boldsymbol{g}_k^0 - \boldsymbol{g}_k\|^2 \leq L_f^2\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\|^2 = L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2. \tag{31}$$

Then, from (31) and $\rho(\boldsymbol{H}) = 1$, we have

$$\|\bar{\boldsymbol{g}}_k^0 - \bar{\boldsymbol{g}}_k\|^2 = \|\boldsymbol{H}(\boldsymbol{g}_k^0 - \boldsymbol{g}_k)\|^2$$
$$\leq \|\boldsymbol{g}_k^0 - \boldsymbol{g}_k\|^2 \leq L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2. \tag{32}$$

From $\nabla\tilde{f}$ is Lipschitz-continuous and (30), we have

$$\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 \leq L_f^2\|\bar{\boldsymbol{x}}_{k+1} - \bar{\boldsymbol{x}}_k\|^2 = \eta^2 L_f^2\|\bar{\boldsymbol{g}}_k\|^2. \tag{33}$$

From [48, Lemma 1.2.3], we know that (7) implies

$$|f_i(y) - f_i(x) - (y-x)^\top\nabla f_i(x)|$$
$$\leq \frac{L_f}{2}\|y-x\|^2 \forall x, y \in \mathbb{R}^d. \tag{34}$$

From (30) and (34), we have

$$\tilde{f}(\bar{\boldsymbol{x}}_{k+1}) - \tilde{f}(\bar{\boldsymbol{x}}_k) \leq -\eta\bar{\boldsymbol{g}}_k^\top\boldsymbol{g}_k^0 + \frac{\eta^2 L_f}{2}\|\bar{\boldsymbol{g}}_k\|^2. \tag{35}$$

2) This step is to show the relation between $U_{1,k+1}$ and $U_{1,k}$. We have

$$U_{1,k+1} = \frac{1}{2}\|\boldsymbol{x}_{k+1}\|_{\boldsymbol{K}}^2 = \frac{1}{2}\|\boldsymbol{x}_k - \eta(\alpha\boldsymbol{L}\hat{\boldsymbol{x}}_k + \beta\boldsymbol{v}_k + \boldsymbol{g}_k)\|_{\boldsymbol{K}}^2$$

$$= \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \eta\alpha\boldsymbol{x}_k^\top\boldsymbol{L}\hat{\boldsymbol{x}}_k + \|\hat{\boldsymbol{x}}_k\|_{\frac{\eta^2\alpha^2}{2}\boldsymbol{L}^2}^2$$

$$- \eta\beta(\boldsymbol{x}_k^\top - \eta\alpha\hat{\boldsymbol{x}}_k^\top\boldsymbol{L})\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k\right)$$

$$+ \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k\right\|_{\frac{\eta^2\beta^2}{2}\boldsymbol{K}}^2$$

$$= \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \eta\alpha\boldsymbol{x}_k^\top\boldsymbol{L}(\boldsymbol{x}_k + \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k) + \|\hat{\boldsymbol{x}}_k\|_{\frac{\eta^2\alpha^2}{2}\boldsymbol{L}^2}^2$$

$$- \eta\beta(\boldsymbol{x}_k^\top - \eta\alpha\hat{\boldsymbol{x}}_k^\top\boldsymbol{L})\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0 + \frac{1}{\beta}\boldsymbol{g}_k - \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0 + \frac{1}{\beta}\boldsymbol{g}_k - \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\frac{\eta^2\beta^2}{2}\boldsymbol{K}}^2$$

$$\leq \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}_k\|_{\frac{\eta\alpha}{2}\boldsymbol{L} - \frac{\eta}{2}\boldsymbol{K}}^2 + \|\hat{\boldsymbol{x}}_k\|_{\frac{3\eta^2\alpha^2}{2}\boldsymbol{L}^2}^2$$

$$+ \frac{\eta}{2}(1 + 3\eta)\|\boldsymbol{g}_k - \boldsymbol{g}_k^0\|^2 + \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_{\frac{\eta\alpha}{2}\boldsymbol{L}}^2$$

$$- \eta\beta(\hat{\boldsymbol{x}}_k + \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \frac{3\eta^2\beta^2}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2$$

$$\leq \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}_k\|_{\frac{\eta\alpha}{2}\boldsymbol{L} - \frac{\eta}{2}\boldsymbol{K}}^2 + \|\hat{\boldsymbol{x}}_k\|_{\frac{3\eta^2\alpha^2}{2}\boldsymbol{L}^2}^2$$

$$+ \frac{\eta}{2}(1 + 3\eta)\|\boldsymbol{g}_k - \boldsymbol{g}_k^0\|^2 + \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_{\frac{\eta}{2}(\alpha\boldsymbol{L} + 2\beta\rho(L)\boldsymbol{K})}^2$$

$$- \eta\beta\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \frac{6\eta^2\beta^2 + \eta\beta\rho^{-1}(L)}{4}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2$$

$$\leq \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \|\boldsymbol{x}_k\|_{\frac{\eta\alpha}{2}\boldsymbol{L} - \frac{\eta}{2}\boldsymbol{K} - \frac{\eta}{2}(1+3\eta)L_f^2\boldsymbol{K}}^2 + \|\hat{\boldsymbol{x}}_k\|_{\frac{3\eta^2\alpha^2}{2}\boldsymbol{L}^2}^2$$

$$- \eta\beta\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right) + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\frac{6\eta^2\beta^2\rho(L) + \eta\beta}{4}\boldsymbol{P}}^2$$

$$+ \frac{\eta}{2}(\alpha + 2\beta)\rho(L)\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|^2 \tag{36}$$

where the second and third equalities hold due to (28c) and (25a), respectively; the first and second inequalities hold due to (24) and $\rho(\boldsymbol{K}) = 1$; and the last inequality holds due to (26c) and (31).

3) This step is to show the relation between $U_{2,k+1}$ and $U_{2,k}$. We have

$$U_{2,k+1} = \frac{1}{2}\left\|\boldsymbol{v}_{k+1} + \frac{1}{\beta}\boldsymbol{g}_{k+1}^0\right\|_{\frac{\alpha+\beta}{\beta}\boldsymbol{P}}^2$$

$$= \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0 + \eta\beta\boldsymbol{L}\hat{\boldsymbol{x}}_k + \frac{1}{\beta}(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)\right\|_{\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}}^2$$

$$= \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}}^2 + \eta(\alpha+\beta)\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \|\hat{\boldsymbol{x}}_k\|_{\frac{\eta^2\beta}{2}(\alpha+\beta)\boldsymbol{L}}^2 + \frac{1}{2\beta^2}\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|_{\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}}^2$$

$$+ \frac{1}{\beta}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)^\top\left(\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}\right)(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)$$

$$+ \eta\hat{\boldsymbol{x}}_k^\top\left(\boldsymbol{K} + \frac{\alpha}{\beta}\boldsymbol{K}\right)(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)$$

$$\leq \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}}^2 + \eta(\alpha+\beta)\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \|\hat{\boldsymbol{x}}\|_{\frac{\eta^2\beta}{2}(\alpha+\beta)\boldsymbol{L} + \frac{\eta^2}{2}\boldsymbol{K}}^2 + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\frac{\eta\beta}{4}\boldsymbol{P}}^2$$

$$+ \|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|_{\frac{(\alpha+\beta)^2}{\eta\beta^5}\boldsymbol{P} + \frac{\alpha+\beta}{2\beta^3}\boldsymbol{P}}^2$$

$$+ \frac{1}{2}\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 + \frac{\eta\alpha}{\beta}\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}(\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0)$$

$$\leq \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P} + \frac{\alpha}{\beta}\boldsymbol{P}}^2 + \eta(\alpha+\beta)\hat{\boldsymbol{x}}_k^\top\boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right)$$

$$+ \|\hat{x}\|^2_{\frac{\eta^2 \beta}{2}(\alpha+\beta)L + \frac{\eta^2}{2}K} + \left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\frac{\eta\beta}{4}P}$$

$$+ c_1\|g_{k+1}^0 - g_k^0\|^2 + \frac{\eta\alpha}{\beta}\hat{x}_k^\top K(g_{k+1}^0 - g_k^0)$$

$$\leq \frac{1}{2}\left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\frac{\alpha+\beta}{\beta}P} + \eta(\alpha+\beta)\hat{x}_k^\top K\left(v_k + \frac{1}{\beta}g_k^0\right)$$

$$+ \|\hat{x}\|^2_{\frac{\eta^2 \beta}{2}(\alpha+\beta)L + \frac{\eta^2}{2}K} + \left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\frac{\eta\beta}{4}P}$$

$$+ \eta^2 c_1 L_f^2 \|\bar{g}_k\|^2 + \frac{\eta\alpha}{\beta}\hat{x}_k^\top K(g_{k+1}^0 - g_k^0) \quad (37)$$

where the second and third equalities hold due to (28d) and (26b), respectively; the first, second, and last inequalities hold due to (24), (26c), and (33), respectively.

4) This step is to show the relation between $U_{3,k+1}$ and $U_{3,k}$. We have

$$U_{3,k+1} = x_{k+1}^\top KP\left(v_{k+1} + \frac{1}{\beta}g_{k+1}^0\right)$$

$$= (x_k - \eta(\alpha L\hat{x}_k + \beta v_k + g_k^0 + g_k - g_k^0))^\top KP(v_k$$

$$+ \frac{1}{\beta}g_k^0 + \eta\beta L\hat{x}_k + \frac{1}{\beta}(g_{k+1}^0 - g_k^0))$$

$$= (x_k^\top KP - \eta(\alpha + \eta\beta^2)\hat{x}_k^\top K)\left(v_k + \frac{1}{\beta}g_k^0\right)$$

$$+ \eta\beta x_k^\top K\hat{x}_k - \|\hat{x}_k\|^2_{\eta^2\alpha\beta L}$$

$$+ \frac{1}{\beta}(x_k^\top KP - \eta\alpha\hat{x}_k^\top K)(g_{k+1}^0 - g_k^0)$$

$$- \eta(\beta v_k + g_k^0 + g_k - g_k^0 - \bar{g}_k)^\top P\left(v_k + \frac{1}{\beta}g_k^0\right)$$

$$- \eta\left(v_k + \frac{1}{\beta}g_k^0\right)^\top PK(g_{k+1}^0 - g_k^0)$$

$$- \eta(g_k - g_k^0)^\top\left(\eta\beta K\hat{x}_k + \frac{1}{\beta}KP(g_{k+1}^0 - g_k^0)\right)$$

$$\leq (x_k^\top KP - \eta\alpha\hat{x}_k^\top K)\left(v_k + \frac{1}{\beta}g_k^0\right) + \|x_k\|^2_{\frac{\eta(\beta+2)}{4}K}$$

$$+ \|\hat{x}_k\|^2_{\eta\beta K + \eta^2(\beta^2 K - \alpha\beta L)} + \frac{\eta}{4}(1 + 6\eta)\|g_k - g_k^0\|^2$$

$$+ \|g_{k+1}^0 - g_k^0\|^2_{\frac{\eta+1}{2\eta\beta^2}P^2 + \frac{1}{4}I_{nd}} - \frac{\eta\alpha}{\beta}\hat{x}_k^\top K(g_{k+1}^0 - g_k^0)$$

$$+ \frac{\eta}{8}\|\bar{g}_k\|^2 - \left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\eta\beta P - 3\eta P^2 - \eta^2 P^2 - \frac{\eta^2\beta^2}{2}I_{np}}$$

$$\leq (x_k^\top KP - \eta\alpha\hat{x}_k^\top K)\left(v_k + \frac{1}{\beta}g_k^0\right) + \|x_k\|^2_{\frac{\eta(\beta+2)}{4}K}$$

$$+ \|\hat{x}_k\|^2_{\eta\beta K + \eta^2(\beta^2 K - \alpha\beta L)} + \frac{\eta}{4}(1 + 6\eta)\|g_k - g_k^0\|^2$$

$$+ c_2\|g_{k+1}^0 - g_k^0\|^2 - \frac{\eta\alpha}{\beta}\hat{x}_k^\top K(g_{k+1}^0 - g_k^0) + \frac{\eta}{8}\|\bar{g}_k\|^2$$

$$- \left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\eta(\beta - 3\rho_2^{-1}(L))P - \eta^2(\rho_2^{-1}(L) + \frac{\beta^2}{2}\rho(L))P}$$

$$\leq x_k^\top KP\left(v_k + \frac{1}{\beta}g_k^0\right) - \eta\alpha\hat{x}_k^\top K\left(v_k + \frac{1}{\beta}g_k^0\right)$$

$$+ \|\hat{x}_k\|^2_{\eta\beta K + \eta^2(\beta^2 K - \alpha\beta L)} + \|x_k\|^2_{\frac{\eta(\beta+2)}{4}K + \frac{\eta}{4}(1+6\eta)L_f^2 K}$$

$$- \frac{\eta\alpha}{\beta}\hat{x}_k^\top K(g_{k+1}^0 - g_k^0) + \left(\eta^2 c_2 L_f^2 + \frac{\eta}{8}\right)\|\bar{g}_k\|^2$$

$$- \left\|v_k + \frac{1}{\beta}g_k^0\right\|^2_{\eta(\beta - 3\rho_2^{-1}(L))P - \eta^2(\rho_2^{-1}(L) + \frac{\beta^2}{2}\rho(L))P} \quad (38)$$

where the second equality holds due to (28c) and (28d); the third equality holds due to (25a), (26b), (29), and $K = I_{nd} - H$; the first inequality holds due (24) and $\rho(K) = 1$; the second inequality holds due to (26c); and the last inequality holds due to (31) and (33).

5) This step is to show the relation between $U_{4,k+1}$ and $U_{4,k}$. We have

$$U_{4,k+1} = n(f(\bar{x}_{k+1}) - f^*) = \tilde{f}(\bar{x}_{k+1}) - nf^*$$

$$= \tilde{f}(\bar{x}_k) - nf^* + \tilde{f}(\bar{x}_{k+1}) - \tilde{f}(\bar{x}_k)$$

$$\leq \tilde{f}(\bar{x}_k) - nf^* - \eta\bar{g}_k^\top g_k^0 + \frac{\eta^2 L_f}{2}\|\bar{g}_k\|^2$$

$$= \tilde{f}(\bar{x}_k) - nf^* - \eta\bar{g}_k^\top g_k^0 + \frac{\eta^2 L_f}{2}\|\bar{g}_k\|^2$$

$$= n(f(\bar{x}_k) - f^*) - \frac{\eta}{2}\bar{g}_k^\top(\bar{g}_k + \bar{g}_k^0 - \bar{g}_k)$$

$$- \frac{\eta}{2}(\bar{g}_k - \bar{g}_k^0 + \bar{g}_k^0)^\top \bar{g}_k^0 + \frac{\eta^2 L_f}{2}\|\bar{g}_k\|^2$$

$$\leq n(f(\bar{x}_k) - f^*) - \frac{\eta}{4}\|\bar{g}_k\|^2 + \frac{\eta}{4}\|\bar{g}_k^0 - \bar{g}_k\|^2 - \frac{\eta}{4}\|\bar{g}_k^0\|^2$$

$$+ \frac{\eta}{4}\|\bar{g}_k^0 - \bar{g}_k\|^2 + \frac{\eta^2 L_f}{2}\|\bar{g}_k\|^2$$

$$= n(f(\bar{x}_k) - f^*) - \frac{\eta}{4}(1 - 2\eta L_f)\|\bar{g}_k\|^2 + \frac{\eta}{2}\|\bar{g}_k^0 - \bar{g}_k\|^2$$

$$- \frac{\eta}{4}\|\bar{g}_k^0\|^2$$

$$\leq n(f(\bar{x}_k) - f^*) - \frac{\eta}{4}(1 - 2\eta L_f)\|\bar{g}_k\|^2 + \|x_k\|^2_{\frac{\eta}{2}L_f^2 K}$$

$$- \frac{\eta}{4}\|\bar{g}_k^0\|^2 \quad (39)$$

where the first inequality holds due to (35); the fourth equality holds due to $\bar{g}_k^\top g_k^0 = g_k^\top H g_k^0 = g_k^\top H H g_k^0 = \bar{g}_k^\top \bar{g}_k^0$; the second inequality holds due to (24); and the last inequality holds due to (32).

6) This step is to show the relation between $\|x_{k+1} - a_{k+1}\|^2$ and $\|x_k - a_k\|^2$. Denote $\mathcal{C}_r(\cdot) = \mathcal{C}(\cdot)/r$, then we have

$$\mathbf{E}_\mathcal{C}[\|x_{k+1} - a_{k+1}\|^2]$$

$$= \mathbf{E}_\mathcal{C}[\|x_{k+1} - x_k + x_k - a_k - \psi q_k\|^2]$$

$$= \mathbf{E}_\mathcal{C}[\|x_{k+1} - x_k + (1 - \psi r)(x_k - a_k)$$

$$+ \psi r(x_k - a_k - \mathcal{C}_r(x_k - a_k))\|^2]$$

$$\leq (1 + c_3^{-1})\mathbf{E}_\mathcal{C}[\|x_{k+1} - x_k\|^2]$$

$$+ (1 + c_3)\mathbf{E}_\mathcal{C}[\|(1 - \psi r)(x_k - a_k)$$

$$+ \psi r(x_k - a_k - \mathcal{C}_r(x_k - a_k))\|^2]$$

$$\leq (1 + c_3^{-1})\mathbf{E}_\mathcal{C}[\|x_{k+1} - x_k\|^2]$$

$$+ (1 + c_3)(1 - \psi r)\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2]$$

$$+ (1 + c_3)\psi r \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k - \boldsymbol{a}_k - \mathcal{C}_r(\boldsymbol{x}_k - \boldsymbol{a}_k)\|^2]$$

$$\leq (1 + c_3^{-1})\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2]$$

$$+ (1 + c_3)(1 - \psi r)\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2]$$

$$+ (1 + c_3)\psi r(1 - \varphi)\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2]$$

$$= (1 + c_3^{-1})\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2]$$

$$+ (1 - c_3 - 2c_3^2)\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2] \quad (40)$$

where the first and second equalities hold due to (28a) and (28e), respectively; the first inequality holds due to (24) and $c_3 > 0$; the second inequality holds due to (24) and $\psi r \in (0, 1]$; and the last inequality holds due to (3).

We have

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 = \eta^2 \|\alpha \boldsymbol{L}\hat{\boldsymbol{x}}_k + \beta \boldsymbol{v}_k + \boldsymbol{g}_k\|^2$$

$$= \eta^2 \|\alpha \boldsymbol{L}(\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k) + \alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{v}_k + \boldsymbol{g}_k^0 + \boldsymbol{g}_k - \boldsymbol{g}_k^0\|^2$$

$$\leq 4\eta^2(\alpha^2 \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_{\boldsymbol{L}^2}^2 + \alpha^2 \|\boldsymbol{x}_k\|_{\boldsymbol{L}^2}^2$$

$$+ \|\beta \boldsymbol{v}_k + \boldsymbol{g}_k^0\|^2 + \|\boldsymbol{g}_k - \boldsymbol{g}_k^0\|^2)$$

$$\leq 4\eta^2 \left(\alpha^2 \rho^2(L) \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|^2 + \|\boldsymbol{x}_k\|_{(\alpha^2 \rho^2(L) + L_f^2)\boldsymbol{K}}^2 \right.$$

$$\left. + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\beta^2 \rho^2(L)\boldsymbol{P}}^2 \right) \quad (41)$$

where the first equality holds due to (28c); the first inequality holds due to (24); and the second inequality holds due to (25b), (26c), and (31).

From (12), (40), and (41), we have

$$\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_{k+1} - \boldsymbol{a}_{k+1}\|^2]$$

$$\leq \mathbf{E}_{\mathcal{C}} \left[(1 - c_4)\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2 + \|\boldsymbol{x}_k\|_{4\eta^2(1+c_3^{-1})(\alpha^2\rho^2(L)+L_f^2)\boldsymbol{K}}^2 \right.$$

$$\left. + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{4\eta^2(1+c_3^{-1})\beta^2\rho^2(L)\boldsymbol{P}}^2 \right]. \quad (42)$$

7) This step is to show the relation between V_{k+1} and V_k. We have

$$\mathbf{E}_{\mathcal{C}}[V_{k+1}]$$

$$\leq \mathbf{E}_{\mathcal{C}} \left[V_k - \|\boldsymbol{x}_k\|_{\eta \boldsymbol{M}_1 - \eta^2 c_6 \boldsymbol{K}}^2 + \|\hat{\boldsymbol{x}}_k\|_{\eta\beta\boldsymbol{K}+\eta^2\boldsymbol{M}_2}^2 \right.$$

$$- \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\eta(\epsilon_3 - \eta\epsilon_4)\boldsymbol{P}}^2 - \eta(\epsilon_5 - \eta\epsilon_6)\|\bar{\boldsymbol{g}}_k\|^2$$

$$\left. - \left(c_4 - \frac{\eta}{2}(\alpha + 2\beta)\rho(L)r_0\right)\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2 - \frac{\eta}{4}\|\bar{\boldsymbol{g}}_k^0\|^2 \right]$$

$$\leq \mathbf{E}_{\mathcal{C}} \left[V_k - \|\boldsymbol{x}_k\|_{\eta(c_5 - \eta c_6)\boldsymbol{K}}^2 + \|\hat{\boldsymbol{x}}_k\|_{\eta(\beta+\eta c_7)\boldsymbol{K}}^2 \right.$$

$$- \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\eta(\epsilon_3 - \eta\epsilon_4)\boldsymbol{P}}^2 - \eta(\epsilon_5 - \eta\epsilon_6)\|\bar{\boldsymbol{g}}_k\|^2$$

$$\left. - \left(c_4 - \frac{\eta}{2}(\alpha + 2\beta)\rho(L)r_0\right)\|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2 - \frac{\eta}{4}\|\bar{\boldsymbol{g}}_k^0\|^2 \right]$$

$$\quad (43)$$

where the first inequality holds due to (12), (36)–(39), and (42); and the second inequality holds due to (25b) and $\beta \leq \alpha$.

For the third term in the right-hand side of (43), from (24) and $\rho(\boldsymbol{K}) = 1$, we have

$$\|\hat{\boldsymbol{x}}_k\|_{\boldsymbol{K}}^2 = \|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k + \boldsymbol{x}_k\|_{\boldsymbol{K}}^2$$

$$\leq 2\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|^2 + 2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2. \quad (44)$$

From (12), (43), and (44), we know that (27) holds. ∎

We are now ready to prove Theorem 1.

We first show that all of the used constants are positive.

From $\alpha = \kappa_1\beta$, $\kappa_1 \geq \frac{9+\kappa_4}{2\rho_2(L)}$, $\kappa_4 > 0$, and $\beta > \kappa_2 \geq \frac{4+5L_f^2}{\kappa_4}$, we have

$$\epsilon_1 = \frac{\kappa_1\beta}{2}\rho_2(L) - \frac{1}{4}(9\beta + 4 + 5L_f^2)$$

$$> \frac{\kappa_1\beta}{2}\rho_2(L) - \frac{1}{4}(9\beta + \kappa_4\beta) \geq 0. \quad (45)$$

From $\beta > \kappa_2 \geq \frac{6}{\rho_2(L)}$, we have

$$\epsilon_3 > 0. \quad (46)$$

From $\alpha = \kappa_1\beta$ and $\beta > \kappa_2 = \max\{\kappa_5, \sqrt{\kappa_6}\}$, we have

$$\epsilon_5 = \frac{1}{8} - \frac{(\kappa_1 + 1)^2 L_f^2}{\beta^3 \rho_2(L)} - \frac{L_f^2}{2\beta^2 \rho_2^2(L)}$$

$$> \frac{1}{8} - \frac{(\kappa_1 + 1)^2 L_f^2}{\kappa_5 \beta^2 \rho_2(L)} - \frac{L_f^2}{2\beta^2 \rho_2^2(L)} > 0. \quad (47)$$

From (45) to (47), and $0 < \eta < \kappa_3$, we have

$$\eta(\epsilon_1 - \eta\epsilon_2) > 0 \quad (48a)$$

$$\eta(\epsilon_3 - \eta\epsilon_4) > 0 \quad (48b)$$

$$\eta(\epsilon_5 - \eta\epsilon_6) > 0 \quad (48c)$$

$$\epsilon_7 - \eta\epsilon_8 - \eta^2\epsilon_9 > 0. \quad (48d)$$

We then show that (13a) and (13b) hold.

From (24), we have

$$V_k \geq \frac{1}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + \frac{1}{2}\left(1 + \frac{\alpha}{\beta}\right)\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P}}^2$$

$$- \frac{\beta}{2\alpha\rho_2(L)}\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 - \frac{\alpha}{2\beta}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P}}^2$$

$$+ n(f(\bar{x}_k) - f^*) + \|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2$$

$$\geq \epsilon_{10}\left(\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{P}}^2\right)$$

$$+ n(f(\bar{x}_k) - f^*) + \|\boldsymbol{x}_k - \boldsymbol{a}_k\|^2 \quad (49a)$$

$$\geq \epsilon_{10}\hat{V}_k \geq 0. \quad (49b)$$

From (27) and (48b)–(48d), we have

$$\mathbf{E}_{\mathcal{C}}[V_{T+1}]$$

$$\leq V_0 - \sum_{k=0}^{T} \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k\|_{\eta(\epsilon_1 - \eta\epsilon_2)\boldsymbol{K}}^2] - \sum_{k=0}^{T} \mathbf{E}_{\mathcal{C}}\left[\frac{\eta}{4}\|\bar{\boldsymbol{g}}_k^0\|^2\right]. \quad (50)$$

From (50), (48a), and (49b), we have

$$\sum_{k=0}^{T} \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + \|\bar{\boldsymbol{g}}_k^0\|^2] \leq \frac{V_0}{\min\{\eta(\epsilon_1 - \eta\epsilon_2), \frac{\eta}{4}\}}$$

which yields (13a).

From (50), (48a), and (49a), we have

$$\mathbf{E}_{\mathcal{C}}[n(f(\bar{x}_T) - f^*)] \leq \mathbf{E}_{\mathcal{C}}[V_T] \leq V_0$$

which yields (13b).

C. Proof of Theorem 2

In this proof, in addition to the notations used in the proof of Theorem 1, we also denote

$$\epsilon = \frac{\epsilon_{12}}{\epsilon_{11}}, \epsilon_{11} = \max\left\{\frac{1}{2} + \frac{\alpha}{\beta}, \frac{\alpha\rho_2(L) + \beta}{2\alpha\rho_2(L)}\right\}$$

$$\epsilon_{12} = \eta \min\left\{\epsilon_1 - \eta\epsilon_2, \epsilon_3 - \eta\epsilon_4, \frac{\nu}{2}, \frac{\epsilon_7}{\eta} - \epsilon_8 - \eta\epsilon_9\right\}.$$

We first show that $\epsilon \in (0, 1)$.

From (48a) to (48d), we have

$$\epsilon_{12} > 0 \text{ and } \epsilon = \frac{\epsilon_{12}}{\epsilon_{11}} > 0. \tag{51}$$

Noting that $\epsilon_{11} \geq \frac{1}{2} + \frac{\alpha}{\beta} \geq \frac{3}{2}$, and $\epsilon_{12} < \epsilon_7 = \frac{\varphi\psi r}{2}(1 + \varphi\psi r) < 1$ due to $\varphi\psi r \in (0, 1)$, we have

$$0 < \epsilon = \frac{\epsilon_{12}}{\epsilon_{11}} < 1. \tag{52}$$

We then show that (14) holds.

From (24), we have

$$V_k \leq \epsilon_{11}\hat{V}_k. \tag{53}$$

From Assumptions 5 and 7 as well as (2), we have that

$$\|\bar{g}_k^0\|^2 = n\|\nabla f(\bar{x}_k)\|^2 \geq 2\nu n(f(\bar{x}_k) - f^*). \tag{54}$$

Then, from (27), (48c), (51), and (53)–(54), we have

$$\mathbf{E}_{\mathcal{C}}[V_{k+1}] \leq \mathbf{E}_{\mathcal{C}}[V_k - \epsilon_{12}\hat{V}_k] \leq \mathbf{E}_{\mathcal{C}}\left[V_k - \frac{\epsilon_{12}}{\epsilon_{11}}V_k\right]. \tag{55}$$

Hence, from (55) and (52), we have

$$\mathbf{E}_{\mathcal{C}}[V_{k+1}] \leq (1 - \epsilon)\mathbf{E}_{\mathcal{C}}[V_k] \leq (1 - \epsilon)^{k+1}V_0$$

which yields (14).

D. Constants Used in Theorem 3

$$\kappa_0 = \frac{1}{\sqrt{r_0 r_1}}, \check{\kappa}_3 = \min\left\{\frac{\epsilon_1}{\epsilon_2}, \frac{\epsilon_3}{\epsilon_4}, \frac{\epsilon_5}{\epsilon_6}, \tau_1, \tau_2\right\}$$

$$\check{\epsilon}_8 = ((\alpha + 2\beta)\rho(L) + 4\beta)(r_0 r_1 + 1)\sigma^2$$

$$\check{\epsilon}_9 = 2(7\alpha^2\rho^2(L) + 2\beta^2 + 1)(r_0 r_1 + 1)\sigma^2$$

$$\epsilon_{10} = \frac{\alpha\rho_2(L) - \beta}{2\alpha\rho_2(L)}$$

$$\check{c}_4 = c_3 + 2c_3^2 - 16\eta^2(1 + c_3^{-1})\alpha^2\rho^2(L)r_0 r_2$$

$$r_1 > 1, r_2 = \frac{r_1}{r_1 - 1}, \tau_0 \in (0, \frac{\epsilon_7}{r_0 r_2})$$

$$\tau_1 = \frac{\sqrt{r_2^2\epsilon_8^2 + 2r_2\epsilon_9(\epsilon_7 - r_0 r_2\tau_0)} - r_2\epsilon_8}{2r_2\epsilon_9}$$

$$\tau_2 = \frac{\sqrt{\check{\epsilon}_8^2 + 4(1 - r_0 r_1\sigma^2)\tau_0\check{\epsilon}_9} - \check{\epsilon}_8}{2\check{\epsilon}_9}$$

$$\tau_3 = \epsilon_7 - r_0 r_2\tau_0 - 2\eta r_2\epsilon_8 - 2\eta^2 r_2\epsilon_9$$

$$\tau_4 = (1 - r_0 r_1\sigma^2)\tau_0 - \eta\check{\epsilon}_8 - \eta^2\check{\epsilon}_9.$$

E. Constants Used in Theorem 4

$$\check{\epsilon} = \frac{\check{\epsilon}_{12}}{\epsilon_{11}}, \check{\epsilon}_{12} = \eta \min\left\{\epsilon_1 - \eta\epsilon_2, \epsilon_3 - \eta\epsilon_4, \frac{\nu}{2}, \frac{\tau_3}{\eta}, \frac{\tau_4}{\eta\tau_0}\right\}.$$

F. Constants Used in Theorem 5

$$\tilde{\kappa}_3 = \min\left\{\frac{\epsilon_1}{\tilde{\epsilon}_2}, \frac{\epsilon_3}{\tilde{\epsilon}_4}, \frac{\epsilon_5}{\epsilon_6}\right\}$$

$$\tilde{\epsilon}_2 = 3L_f^2 + 2\beta^2 + 1 + 3\alpha^2\rho^2(L)$$

$$\tilde{\epsilon}_4 = 2\beta^2\rho(L) + \rho_2^{-1}(L)$$

$$\tilde{\epsilon}_8 = \frac{1}{2}(\alpha + 2\beta)\rho(L) + 2\beta$$

$$c_8 = \eta(\tilde{\epsilon}_8 + 2c_7\eta)n\tilde{d}^2 C s_0^2.$$

G. Constants Used in Theorem 6

$$\tilde{\epsilon} \in (0, \min\{\tilde{\epsilon}_0, 1 - \gamma^2\}), \tilde{\epsilon}_0 = \frac{\tilde{\epsilon}_{12}}{\epsilon_{11}}$$

$$\tilde{\epsilon}_{12} = \eta \min\left\{\epsilon_1 - \eta\tilde{\epsilon}_2, \epsilon_3 - \eta\tilde{\epsilon}_4, \frac{\nu}{2}\right\}.$$

H. Constants Used in Theorem 7

$$\hat{\kappa}_3 = \min\left\{\frac{\hat{\epsilon}_1}{2\hat{\epsilon}_2}, \frac{\epsilon_5}{\epsilon_6}, \sqrt{\kappa_{15}}\right\}, \gamma \in [\max\{\sqrt{\kappa_{11}}, \sqrt{\kappa_{12}}\}, 1)$$

$$s_0 \geq \max\{\sqrt{\kappa_8/\kappa_7}, \max_{i\in[n]}\|x_{i,0}\|\}, \kappa_7 > \frac{\kappa_{10}(1 - \varphi)^2}{\kappa_9}$$

$$\kappa_8 = \frac{1}{2}\|x_0\|_K^2 + \frac{1}{2}\left\|v_0 + \frac{1}{\beta}g_0^0\right\|_{\frac{\alpha+\beta}{\beta}P}^2$$

$$+ x_0^\top KP\left(v_0 + \frac{1}{\beta}g_0^0\right) + \frac{1}{2\nu}\|\bar{g}_0^0\|^2$$

$$\kappa_9 = \min\left\{\frac{\hat{\epsilon}_1}{2\epsilon_{11}}, \frac{\nu}{2\epsilon_{11}}\right\}, \kappa_{10} = \left(\tilde{\epsilon}_8 + \frac{c_7\hat{\epsilon}_1}{\hat{\epsilon}_2}\right)n\tilde{d}^2$$

$$\kappa_{11} = 1 - \eta\kappa_9 + \frac{\eta\kappa_{10}(1 - \varphi)^2}{\kappa_7}$$

$$\kappa_{12} = (1 + \varphi + \eta^2\kappa_{13})(1 - \varphi)^2 + \frac{\eta^2\kappa_{14}}{\epsilon_{10}}\kappa_7$$

$$\kappa_{13} = 4(1 + \varphi^{-1})\hat{d}^2 n\tilde{d}^2\alpha^2\rho^2(L)$$

$$\kappa_{14} = 4(1 + \varphi^{-1})\hat{d}^2(\alpha^2\rho^2(L) + L_f^2)$$

$$\kappa_{15} = \frac{(\varphi + \varphi^2 - \varphi^3)\epsilon_{10}}{\kappa_{13}(1 - \varphi)^2\epsilon_{10} + \kappa_{14}\kappa_7}$$

$$\hat{\epsilon}_0 = \frac{\hat{\epsilon}_{12}}{\epsilon_{11}}, \hat{\epsilon}_1 = \min\{\epsilon_1, \epsilon_3\}, \hat{\epsilon}_2 = \max\{\tilde{\epsilon}_2, \tilde{\epsilon}_4\}$$

$$\hat{\epsilon}_{12} = \eta \min \left\{ \hat{\epsilon}_1 - \eta \hat{\epsilon}_2, \frac{\nu}{2} \right\}, c_9 = \eta(\tilde{\epsilon}_8 + 2c_7\eta)n\tilde{d}^2.$$

REFERENCES

[1] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013.

[2] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 707–724, Aug. 2011.

[3] H.-T. Wai, T.-H. Chang, and A. Scaglione, "A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 3546–3550.

[4] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[5] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-PDA: The proximal primal–dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1529–1538.

[6] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of gradient algorithms over networks," in *Proc. Annu. Allerton Conf. Commun., Control, Comput.*, 2018, pp. 359–365.

[7] D. Hajinezhad and M. Hong, "Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization," *Math. Program.*, vol. 176, no. 1/2, pp. 207–245, 2019.

[8] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5912–5928, Nov. 2019.

[9] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization," *IEEE Trans. Autom. Control*, vol. 67, no. 8, pp. 4194–4201, Aug. 2022.

[10] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, "Distributed learning in the nonconvex world: From batch data to streaming and beyond," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 26–38, May 2020.

[11] R. Xin, U. A. Khan, and S. Kar, "An improved convergence analysis for decentralized online stochastic non-convex optimization," *IEEE Trans. Signal Process.*, vol. 69, pp. 1842–1858, 2021.

[12] Y. Lu and C. De Sa, "Moniqua: Modulo quantized communication in decentralized SGD," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6415–6425.

[13] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, "On biased compression for distributed learning," 2020, *arXiv:2002.12410*.

[14] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.

[15] S. Horváth, D. Kovalev, K. Mishchenko, P. Richtárik, and S. Stich, "Stochastic distributed learning with gradient quantization and double-variance reduction," *Optim. Methods Softw.*, pp. 1–16, 2022.

[16] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3478–3487.

[17] 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.

[18] X. Liu, Y. Li, R. Wang, J. Tang, and M. Yan, "Linear convergent decentralized optimization with compression," in *Proc. Int. Conf. Learn. Representations*, 2021.

[19] 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.

[20] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "Distributed subgradient methods and quantization effects," in *Proc. IEEE Conf. Decis. Control*, 2008, pp. 4177–4184.

[21] S. Zhu, M. Hong, and B. Chen, "Quantized consensus ADMM for multi-agent distributed optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4134–4138.

[22] D. Yuan, S. Xu, H. Zhao, and L. Rong, "Distributed dual averaging method for multi-agent optimization with quantized communication," *Syst. Control Lett.*, vol. 61, no. 11, pp. 1053–1061, 2012.

[23] 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.

[24] T. T. Doan, S. T. Maguluri, and J. Romberg, "Fast convergence rates of distributed subgradient methods with adaptive quantization," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2191–2205, May 2021.

[25] P. 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.

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

[27] S. Magnússon, H. Shokri-Ghadikolaei, and N. Li, "On maintaining linear convergence of distributed learning and optimization under limited communication," *IEEE Trans. Signal Process.*, vol. 68, pp. 6101–6116, 2020.

[28] 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.

[29] 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, doi: 10.1109/TAC.2022.3219289.

[30] J. Lei, P. Yi, G. Shi, and B. D. Anderson, "Distributed algorithms with finite data rates that solve linear equations," *SIAM J. Optim.*, vol. 30, no. 2, pp. 1191–1222, 2020.

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

[32] S. Khirirat, S. Magnússon, and M. Johansson, "Compressed gradient methods with Hessian-aided error compensation," *IEEE Trans. Signal Process.*, vol. 69, pp. 998–1011, 2021.

[33] J. Zhang, K. You, and L. Xie, "Innovation compression for communication-efficient distributed optimization with linear convergence," 2021, *arXiv:2105.06697*.

[34] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7663–7673.

[35] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *Proc. Int. Conf. Learn. Representations*, 2020.

[36] 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.

[37] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization," in *Proc. IEEE Conf. Decis. Control*, 2020, pp. 3449–3456.

[38] N. Singh, D. Data, J. George, and S. Diggavi, "SQuARM-SGD: Communication-efficient momentum SGD for decentralized optimization," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 954–969, Sep. 2021.

[39] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, "Robust and communication-efficient collaborative learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8386–8397.

[40] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2016, pp. 795–811.

[41] Y. Tang, J. Zhang, and N. Li, "Distributed zero-order algorithms for nonconvex multi-agent optimization," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 269–281, Mar. 2021.

[42] Z. Li and J. Li, "A simple proximal stochastic gradient method for nonsmooth nonconvex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5569–5579.

[43] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.

[44] C. Chen, J. Zhang, L. Shen, P. Zhao, and Z. Luo, "Communication efficient primal–dual algorithm for nonconvex nonsmooth distributed optimization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1594–1602.

[45] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Communication compression for distributed nonconvex optimization," 2022, *arXiv:2201.03930v3*.

[46] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[47] X. Yi, L. Yao, T. Yang, J. George, and K. H. Johansson, "Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 3397–3402.

[48] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Berlin, Germany: Springer, 2018.

Xinlei Yi received the B.S. and M.S. degrees in mathematics from the China University of Geoscience, Wuhan, China, and Fudan University, Shanghai, China, in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2020.

He is a Postdoc with the Massachusetts Institute of Technology, Cambridge, MA, USA. He was a Postdoc with the KTH Royal Institute of Technology until 2022. His doctoral thesis was selected as one of the four finalists for the 2021 European Systems and Control Ph.D. Thesis Award. His current research interests include online optimization, distributed optimization, and event-triggered control.

Shengjun Zhang received the B.Eng. degree in automation of honors program from China Agricultural University, Beijing, China, in 2014, the M.S. degree in electrical engineering from New York University, New York City, NY, USA, in 2017, and the Ph.D. degree in electrical engineering from the Department of Electrical Engineering, College of Engineering, University of North Texas, Denton, TX, USA, in 2022.

He is currently with AlgoRhythm, Inc., Wuhan, China. His current research interests include optimization and applications in artificial intelligence.
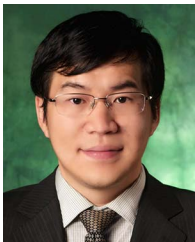
Tao Yang (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Washington State University, Pullman, WA, USA, in 2012.

Between August 2012 and August 2014, he was an ACCESS Postdoctoral Researcher with the ACCESS Linnaeus Center, KTH Royal Institute of Technology, Sweden. He then joined the Pacific Northwest National Laboratory as a Postdoc, and was promoted to Scientist/Engineer II in 2015. He was an Assistant Professor with the Department of Electrical Engineering, University of North Texas, Denton, TX, USA, from 2016 to 2019. He is a Professor with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. His research interests include industrial artificial intelligence, integrated optimization and control, distributed control and optimization with applications to process industries, cyber physical systems, networked control systems, and multiagent systems.

Dr. Yang was a recipient of the Ralph E. Powe Junior Faculty Enhancement Award and Best Student Paper award (as an advisor) of the 14th IEEE International Conference on Control and Automation in 2018. He is an Associate Editor for IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *IEEE/CAA Journal of Automatica Sinica.* He currently is a member of the Technical Committee on Nonlinear Systems and Control, the Technical Committee on Networks and Communication Systems, and the Technical Committee on Smart Grids of the IEEE Control Systems Society, a member of the IEEE Control Systems Society Conference Editorial Board, and a member of the IFAC Technical Committee on Large Scale Complex Systems.

Tianyou Chai (Life Fellow, IEEE) received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985.

He became a Professor with Northeastern University in 1988. He is the Founder and Director of the Center of Automation, which became a National Engineering and Technology Research Center and a State Key Laboratory. He is a member of Chinese Academy of Engineering and IFAC Fellow. He was the Director of the Department of Information Science of National Natural Science Foundation of China from 2010 to 2018. He has authored/coauthored 230 peer-reviewed international journal papers. His paper titled "Hybrid intelligent control for optimal operation of shaft furnace roasting process" was selected as one of three best papers for the Control Engineering Practice Paper Prize for 2011–2013. He has developed control technologies with applications to various industrial processes. His current research interests include modeling, control, optimization, and integrated automation of complex industrial processes.

Dr. Chai was a recipient of five prestigious awards of National Natural Science, National Science and Technology Progress, and National Technological Innovation for his contributions, the 2007 Industry Award for Excellence in Transitional Control Research from IEEE Multiple Conference on Systems and Control, and the 2017 Wook Hyun Kwon Education Award from Asian Control Association.

Karl Henrik Johansson (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1992 and 1997, resptively.

He is a Professor with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. He has held visiting positions with UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests include networked control systems, cyber-physical systems, and applications in transportation, energy, and automation. He has served on the IEEE Control Systems Society Board of Governors, the IFAC Executive Board, and the European Control Association Council.

Dr. Johansson was a recipient of several best paper awards and other distinctions from IEEE and ACM. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He was also a recipient of the Future Research Leader Award from the Swedish Foundation for Strategic Research and the Triennial Young Author Prize from IFAC. He is Fellow of the Royal Swedish Academy of Engineering Sciences, and he is the IEEE Control Systems Society Distinguished Lecturer.