Automatica 177 (2025) 112286

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Compressed gradient tracking algorithms for distributed nonconvex optimization☆

Lei Xu [a,b], Xinlei Yi [c,d], Guanghui Wen [e], Yang Shi [b], Karl H. Johansson [f], Tao Yang [a,*]

[a] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China
[b] Department of Mechanical Engineering, University of Victoria, Victoria, BC V8W 2Y2, Canada
[c] College of Electronics and Information Engineering, Tongji University, Shanghai, 201804, China
[d] Lab for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[e] Department of Systems Science, School of Mathematics, Southeast University, Nanjing 210096, China
[f] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 10044, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 31 October 2023
Received in revised form 26 December 2024
Accepted 10 February 2025
Available online 9 April 2025

Keywords:
Communication compression
Gradient tracking algorithm
Linear convergence
Nonconvex optimization
Polyak–Łojasiewicz condition
Sublinear convergence

ABSTRACT

In this paper, we study the distributed nonconvex optimization problem, aiming to minimize the average value of the local nonconvex cost functions using local information exchange. To reduce the communication overhead, we introduce three general classes of compressors, i.e., compressors with bounded relative compression error, compressors with globally bounded absolute compression error, and compressors with locally bounded absolute compression error. By integrating them, respectively, with the distributed gradient tracking algorithm, we then propose three corresponding compressed distributed nonconvex optimization algorithms. Motivated by the state-of-the-art BEER algorithm proposed in Zhao et al. (2022), which is an efficient compressed algorithm integrating gradient tracking with biased and contractive compressors, our first proposed algorithm extends this algorithm to accommodate both biased and non-contractive compressors For each algorithm, we design a novel Lyapunov function to demonstrate its sublinear convergence to a stationary point if the local cost functions are smooth. Furthermore, when the global cost function satisfies the Polyak–Łojasiewicz (P–Ł) condition, we show that our proposed algorithms linearly converge to a global optimal point. It is worth noting that, for compressors with bounded relative compression error and globally bounded absolute compression error, our proposed algorithms' parameters do not require prior knowledge of the P–Ł constant.

1. Introduction

In recent years, distributed optimization has received considerable attention and has been applied in power systems, sensor networks, and machine learning, just to name a few (Du et al., 2018; Guo et al., 2016; Sayed, 2014; Zhang & Wang, 2019). Various distributed optimization algorithms have been proposed; see, e.g., survey papers (Nedić & Liu, 2018; Yang et al., 2019) and references therein. Early work (Nedić & Ozdaglar, 2009) proposed a distributed (sub)gradient descent (DGD) algorithm that requires a diminishing step size. In order to speed up the convergence rate, the accelerated distributed optimization algorithms with a fixed step size have been proposed, such as EXTRA algorithm (Shi et al., 2015), distributed proportional-integral (DPI) algorithm (Kia et al., 2015; Wang & Elia, 2010), and distributed gradient tracking (DGT) algorithm (Nedić et al., 2017; Qu & Li, 2017; Xu et al., 2015).

Distributed optimization algorithms involve data spread across multiple agents, which require each agent to communicate with its neighboring agents. However, one of the main challenges is the communication bottleneck, which can occur owing to limited channel bandwidth or communication power. Communication compression, which encompasses techniques like quantization and sparsification, can reduce the required communication capacity, see recent survey papers (Cao et al., 2023; Shi et al., 2020). For distributed convex optimization problems, various compressed distributed optimization algorithms have been developed. For example, based on the DGD algorithm, Doan et al. (2020), Yi and Hong (2014), Zhang et al. (2018) developed a quantized gradient algorithm using a uniform quantizer, a random quantizer, and the sign of the relative state, respectively. Alistarh et al. (2017), Horváth et al. (2023) used a distributed stochastic

☆ This paper was recommended for publication in revised form by Associate Editor Luca Schenato under the direction of Editor Christos G. Cassandras. The material in this paper was not presented at any conference.

* Corresponding author.
E-mail addresses: 2010345@stu.neu.edu.cn (L. Xu), xinleiyi@mit.edu (X. Yi), wenguanghui@gmail.com (G. Wen), yshi@uvic.ca (Y. Shi), kallej@kth.se (K.H. Johansson), yangtao@mail.neu.edu.cn (T. Yang).

https://doi.org/10.1016/j.automatica.2025.112286
0005-1098/

gradient descent (DSGD) algorithm with gradient quantization and encoding/variance reduction to design a compressed DSGD algorithm. Liao et al. (2022), Ma et al. (2021), Xiong et al. (2022) developed quantized/compressed DGT algorithms by integrating the gradient tracking method with a uniform quantizer and compressors that have a bounded relative compression error, respectively. Yuan et al. (2012) proposed a distributed averaging method using random quantization.

Previous studies have focused on distributed convex optimization. However, many applications such as distributed learning (Omidshafiei et al., 2017), distributed clustering (Forero et al., 2011), the cost functions are usually nonconvex, see, e.g., Hong et al. (2017), Yi et al. (2021). Consequently, several studies have proposed compressed distributed nonconvex optimization algorithms. For example, Reisizadeh et al. (2019), Taheri et al. (2020) proposed compressed DSGD algorithms that utilize exact and random quantization, respectively. Xu et al. (2022) proposed two quantized distributed algorithms by integrating a uniform quantizer with the DGT and DPI algorithms , respectively. Liao et al. (2023) proposed a compressed DGT algorithm that utilizes robust compressors. Furthermore, Yi et al. (2023) developed compressed distributed primal–dual algorithms, which used several general compressors. In this paper, we investigate another distributed optimization algorithm equipped with these general compressors, i.e., DGT algorithm. As explained by Koloskova et al. (2021), EXTRA (Shi et al., 2015) and distributed primal–dual (Alghunaim & Sayed, 2020) algorithms typically demand a noiseless setting. In contrast, the DGT algorithm is able to tolerate stochastic noise (Di Lorenzo & Scutari, 2016; Nedić et al., 2017), rendering it suitable for solving nonconvex optimization problems, especially in machine learning, see, e.g., Lin et al. (2021), Yuan et al. (2021). This motivates us to consider the gradient tracking methods with communication compression for the distributed nonconvex optimization problem.

The main contributions of this paper are:

- For the compressors with bounded relative compression error, which include the commonly considered unbiased and biased but contractive compressors, we design a compressed DGT algorithm (Algorithm 1). For smooth local cost functions, we design an appropriate Lyapunov function in Theorem 1 to show that the proposed algorithm sublinearly converges to a stationary point. Moreover, if the global cost function satisfies the Polyak–Łojasiewicz (P–Ł) condition, which is weaker than the standard strong convexity condition and the global minimizer is not necessarily unique, we establish in Theorem 2 that the proposed algorithm linearly converges to a global optimal point.
- We propose an error feedback based compressed gradient tracking algorithm (Algorithm 2) to improve the algorithm's efficiency for biased compression methods. Moreover, we redesign a Lyapunov function in Theorem 3 to accommodate the introduced error feedback variables, and utilize it to establish convergence results without (Theorem 3) and with (Theorem 4) the P–Ł condition.
- For the compressors with bounded absolute compression error, which includes the commonly considered unbiased compressors with bounded variance, we develop a compressed DGT algorithm (Algorithm 3), and redesign a Lyapunov function in Theorem 5. For compressors with globally bounded absolute compression error, we present the convergence results without (Theorem 5) and with (Theorem 6) the P–Ł condition, which are similar to Theorems 1 and 2.
- For compressors with locally bounded absolute compression error, we redesign a Lyapunov function in Theorem 7, to demonstrate that the proposed Algorithm 3 linearly converge to a global optimal point under the P–Ł condition.

Note that for perfect communication in DGT nonconvex optimization algorithms under the P–Ł condition, Tang et al. (2021), Xin et al. (2021) constructed systems of linear inequalities to analyze the convergence of the algorithms. Nonetheless, the prior knowledge of the P–Ł constant is required to determine their proposed algorithms' parameters. To avoid the need for the P–Ł constant in determining algorithm parameters, this paper employs the Lyapunov method to analyze the convergence of the proposed algorithms. This is a significant property since determining the P–Ł constant can be a challenging task. Moreover, for the perfect communication scenario, Laypunov analysis has been developed for DGT in both the convex (Notarnicola et al., 2023) and nonconvex cases (Carnevale & Notarstefano, 2022). The Lyapunov analysis method is simpler than other methods, which typically require constructing systems of linear inequalities, employing Lyapunov-like arguments, or utilizing control tools, see, e.g., Qu and Li (2017), Tang et al. (2021), Varagnolo et al. (2015), Xin et al. (2021), Xu et al. (2017). In Zhao et al. (2022), the authors also utilized the Lyapunov method to analyze the convergence of the proposed BEER algorithm. However, the compressors with bounded relative compression error that we considered are more general than the biased but contractive compressors used in Zhao et al. (2022). Furthermore, we also consider compressors with globally and locally bounded absolute compression error. To the best of our knowledge, this paper is the first to avoid using the P–Ł constant in the context of the DGT nonconvex optimization algorithm under the P–Ł condition. This work differs from (Yi et al., 2023) in two key aspects: (i) This paper is built upon the DGT algorithm to design compressed distributed nonconvex optimization algorithms, whereas (Yi et al., 2023) is based on the distributed primal–dual algorithm. The primary distinction between these compressed algorithms lies in the fact that the gradient tracking algorithm is better suited for the nonconvex setting. In the numerical simulation section of the online version (Xu et al., 2023), we showcase that the proposed algorithms achieve faster convergence when compared to the compressed algorithms proposed in Yi et al. (2023). (ii) In this paper, for each compressed algorithm, we developed a novel Lyapunov function to analyze the convergence of the proposed algorithm. This Lyapunov function differs from the one proposed in Yi et al. (2023).

The remainder of the paper is organized as follows. Section 2 presents the problem formulation. In Sections 3 and 4, we introduce three compressed gradient tracking algorithms and conduct analyses for compressors involving bounded relative compression error, as well as globally and locally bounded absolute compression errors, respectively. Finally, concluding remarks are offered in Section 5. All the proofs of theorems are given in the online version (Xu et al., 2023) due to space limitations.

Notation. Let $\mathbf{1}_n$ (or $\mathbf{0}_n$) be the $n\times 1$ vector with all ones (or zeros), and $\mathbf{I}_n$ be the n-dimensional identity matrix. $\mathrm{col}(Z_1,\ldots,Z_n)$ is the concatenated column vector of vectors $Z_i\in\mathbb{R}^d$. $\|\cdot\|$ is the Euclidean vector norm or spectral matrix norm. For a column vector $X=(X_1,\ldots,X_m)$, $\|X\|_\infty=\max_{1\le i\le m}|X_i|$. For a positive semi-definite matrix $\mathcal{M}$, $\rho(\mathcal{M})$ is the spectral radius. The minimum integer less than or equal to c is denoted by $\lfloor c\rfloor$. $\mathrm{sign}(c)$ and $|c|$ are the element–wise sign and absolute value, respectively. Given any differentiable function F, ∇F is the gradient of F. $A\otimes B$ represents the Kronecker product of matrices A and B. $A\preceq B$ if all entries of matrix $A-B$ are not greater than zero, and $A\succ 0$ if all entries of matrix A that are greater than zero. $\mathbb{Z}^+$ denotes the set of positive integers.

2. Problem formulation

Consider a group of n agents over a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed edges. A directed path from agent i_1 to agent i_k is a sequence of agents $\{i_1, \ldots, i_k\}$ such that $(i_j, i_{j+1}) \in \mathcal{E}$, $j = 1, \ldots, k-1$. A directed graph is strongly connected if there exists a path between any pair of distinct agents.

Assume that each agent has a private differentiable local cost function $F_i : \mathbb{R}^d \to \mathbb{R}$, the optimal set $\mathbb{X}^\star = argmin_{X\in\mathbb{R}^d} F(X)$ is nonempty and $F^\star = min_{X\in\mathbb{R}^d} F(X) > -\infty$. The objective is to find an optimizer $X^\star$ to minimize the average of all local cost functions $F(X) = \frac{1}{n}\sum_{i=1}^n F_i(X)$, that is,

$$\min_{X\in\mathbb{R}^d} F(X) = \min_{X\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n F_i(X). \tag{1}$$

Throughout this paper, we make the following assumptions.

Assumption 1. The directed graph $\mathcal{G}$ is strongly connected and permits a nonnegative doubly stochastic weight matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$, where $w_{ii} > 0$, for all $i \in \mathcal{V}$, and $w_{ij} > 0$ if and only if agent i can receive information from agent j, otherwise $w_{ij} = 0$. Moreover, $W\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{1}_n^T W = \mathbf{1}_n^T$, ensuring that W is doubly stochastic.

Assumption 2. Each local cost function $F_i(X)$ is smooth with constant $L_f > 0$, i.e.,

$$\|\nabla F_i(X) - \nabla F_i(Y)\| \le L_f \|X - Y\|, \ \forall X, Y \in \mathbb{R}^d. \tag{2}$$

Assumption 3. The global cost function $F(X)$ satisfies the Polyak–Łojasiewicz (P–Ł) condition with $\nu > 0$, i.e.,

$$\frac{1}{2}\|\nabla F(X)\|^2 \ge \nu(F(X) - F^\star), \ \forall X \in \mathbb{R}^d. \tag{3}$$

Assumptions 1 and 2 are common in the literature, e.g., Nedić and Liu (2018), Yang et al. (2019). Assumption 3 does not imply convexity of the global cost function, but it ensures that all stationary points are global optima.

Motivated by scenarios where the communication channel often has limited bandwidth, we propose three compressed distributed nonconvex (without and with the P–Ł condition) optimization algorithms utilizing compressors with bounded relative compression error (Assumption 4), globally bounded absolute compression error (Assumption 5), and locally bounded absolute compression error (Assumption 6), respectively, in the subsequent sections.

3. Compressed distributed nonconvex algorithms: Bounded relative compression error

In this section, we introduce a compression operator with bounded relative compression error. In Section 3.1, we propose a compressed DGT algorithm, and present the convergence results. Additionally, in Section 3.2, we extend the compressed algorithm to an error feedback version for biased compressors, and present the convergence result.

Assumption 4 (*Liao et al., 2022; Yi et al., 2023*)**.** The compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$, adheres to the condition:

$$\mathbf{E}_{\mathcal{C}}[\|\frac{\mathcal{C}(X)}{r} - X\|^2] \le (1-\psi)\|X\|^2, \ \forall X \in \mathbb{R}^d, \tag{4}$$

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

This condition implies that

$$\mathbf{E}_{\mathcal{C}}[\|\mathcal{C}(X) - X\|^2] \le C\|X\|^2, \ \forall X \in \mathbb{R}^d, \tag{5}$$

where $C = 2r^2(1-\psi) + 2(1-r)^2$.

Assumption 4 encompasses various compression operators commonly used in the literature, such as norm-sign compression operators, random quantization, and sparsification (Liao et al., 2022; Yi et al., 2023). It represents a broader class of compressors utilized in distributed optimization algorithms.

3.1. Compressed gradient tracking algorithm: Bounded relative compression error

In this section, we propose the compressed DGT algorithm (Algorithm 1), which is the same as the C-GT proposed in Liao et al. (2022). In Liao et al. (2022), the authors constructed systems of linear inequalities to demonstrate the linear convergence of the proposed algorithm under the strongly convex case. In this section, we demonstrate that the proposed Algorithm 1 exhibits sublinear convergence in the general nonconvex case (Theorem 1) and linear convergence when the P–Ł condition is satisfied (Theorem 2). Furthermore, we design a Lyapunov function for analyzing the convergence of the proposed algorithm. Benefiting from this Lyapunov function, in Theorem 2, we are able to design the algorithm parameters without prior knowledge of the P–Ł constant. This features a unique aspect that separates our work from the existing DGT nonconvex optimization results (Tang et al., 2021; Xin et al., 2021) under the P–Ł condition.

Algorithm 1

For each agent $i \in \mathcal{V}$.
Initialization:
$X_i(0) \in \mathbb{R}^d$, $Y_i(0) = \nabla F_i(X_i(0))$, $A_i(0) = B_i(0) = C_i(0) = D_i(0) = \mathbf{0}_d$, $Q_i^X(0) = \mathcal{C}(X_i(0))$, and $Q_i^Y(0) = \mathcal{C}(Y_i(0))$.
Communication:
Transmit $Q_i^X(k)$ and $Q_i^Y(k)$ to its out-neighbors and receive $Q_j^X(k)$ and $Q_j^Y(k)$ from its in-neighbors.
Update Rule:

$$A_i(k+1) = A_i(k) + \varphi_X Q_i^X(k), \tag{6a}$$

$$B_i(k+1) = B_i(k) + \varphi_X(Q_i^X(k) - \sum_{j=1}^n W_{ij}Q_j^X(k)), \tag{6b}$$

$$C_i(k+1) = C_i(k) + \varphi_Y Q_i^Y(k), \tag{6c}$$

$$D_i(k+1) = D_i(k) + \varphi_Y(Q_i^Y(k) - \sum_{j=1}^n W_{ij}Q_j^Y(k)), \tag{6d}$$

$$X_i(k+1) = X_i(k) - \gamma[B_i(k) + Q_i^X(k) - \sum_{j=1}^n W_{ij}Q_j^X(k)] - \eta Y_i(k), \tag{6e}$$

$$Y_i(k+1) = Y_i(k) - \gamma[D_i(k) + Q_i^Y(k) - \sum_{j=1}^n W_{ij}Q_j^Y(k)] + \nabla F_i(X_i(k+1)) - \nabla F_i(X_i(k)), \tag{6f}$$

$$Q_i^X(k+1) = \mathcal{C}(X_i(k+1) - A_i(k+1)), \tag{6g}$$

$$Q_i^Y(k+1) = \mathcal{C}(Y_i(k+1) - C_i(k+1)), \tag{6h}$$

where γ, η, φ_X, and φ_Y are positive parameters.

Remark 1. The main difference between the proposed Algorithm 1 and BEER proposed in Zhao et al. (2022) lies in the introduction of φ_X and φ_Y to correct the error caused by the

r-scaling of the compression operator. It is easy to verify that Algorithm 1 reduces to BEER when $\varphi_X = \varphi_Y = 1$. Due to the introduction of φ_X and φ_Y, Algorithm 1 is suitable for a broader range of compression operators, including (4). Note that the compressors in (4) are equivalent to the compressors used in Zhao et al. (2022) when $r = 1$. However, the introduction of φ_X and φ_Y, along with $r \neq 1$, makes it challenging to straightforwardly extend the proof techniques used in BEER. More specifically, when analyzing the convergence of compression errors, we must construct inequalities that satisfy the properties of the compressors we consider. This requires frequent use of inequality shrinking techniques, and if the increased conservativeness from these shrinkings is not carefully managed, it may prevent us from proving algorithm convergence. These factors present significant challenges in our analysis.

We denote $\boldsymbol{X} = \mathrm{col}(X_1, \ldots, X_n)$, $\boldsymbol{Y} = \mathrm{col}(Y_1, \ldots, Y_n)$, $\boldsymbol{A} = \mathrm{col}(A_1, \ldots, A_n)$, $\boldsymbol{C} = \mathrm{col}(C_1, \ldots, C_n)$, $\bar{X}(k) = \frac{1}{n}(\mathbf{1}_n^T \otimes \mathbf{I}_d)\boldsymbol{X}(k)$, $\bar{\boldsymbol{X}}(k) = \mathbf{1}_n \otimes \bar{X}(k)$, $\boldsymbol{H} = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^T \otimes \mathbf{I}_d)$, $\bar{\boldsymbol{Y}}(k) = \boldsymbol{H}\boldsymbol{Y}(k)$.

To analyze the convergence of Algorithm 1, we consider the following Lyapunov candidate function

$$U(k) = V(k) + n(F(\bar{X}(k)) - F^\star), \tag{7}$$

where $V(k) = \|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + \phi\|\boldsymbol{Y}(k) - \bar{\boldsymbol{Y}}(k)\|^2 + \|\boldsymbol{X}(k) - \boldsymbol{A}(k)\|^2 + \|\boldsymbol{Y}(k) - \boldsymbol{C}(k)\|^2$, and $\phi = \frac{(1-\sigma)^2}{320L_f^2}$ with $\sigma \in (0, 1)$ being the spectral norm of $W - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$.

Note that the designed Lyapunov candidate function (7) incorporates several nonnegative error terms: the consensus error term $\|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2$, gradient tracking error term $\|\boldsymbol{Y}(k) - \bar{\boldsymbol{Y}}(k)\|^2$, compression error terms $\|\boldsymbol{X}(k) - \boldsymbol{A}(k)\|^2$ and $\|\boldsymbol{Y}(k) - \boldsymbol{C}(k)\|^2$, and the optimal error term $n(F(\bar{X}(k)) - F^\star)$. The weight parameter ϕ is instrumental in fine-tuning the values of the respective terms within the designed Lyapunov candidate function (7), thereby ensuring the convergence of the proposed algorithms.

We are now ready to present the convergence results of Algorithm 1.

Theorem 1. *Suppose that Assumption 1, 2, and 4 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 1 with algorithm parameters φ_X, $\varphi_Y \in (0, \frac{1}{r})$, η and γ such that*

$$\begin{aligned}
\eta \in &(0, \min\{\frac{(1-\sigma)^2\gamma}{40L_f}, \frac{0.4(1-\sigma)\gamma}{L_f^2}, \frac{(1-\sigma)^2}{80L_f}\sqrt{\frac{\gamma}{1+c_1^{-1}}},\\
&\frac{9}{40(4(1+c_1^{-1})+5(1+c_2^{-1}))}, \frac{1}{2L_f}, \gamma\}),\\
\gamma \in &(0, \Pi := \min\{\frac{1-\sigma}{160(1+c_1^{-1})}, \frac{1-\sigma}{40000(1+c_2^{-1})L_f^2},\\
&\frac{c_1(1-\sigma)}{40C}, \frac{c_1}{8\sqrt{C}}, \frac{c_1}{10L_f\sqrt{C(1+c_2^{-1})}}, \frac{c_2L_f^2}{C},\\
&\frac{c_2}{10\sqrt{C}}\}),
\end{aligned} \tag{8}$$

where $c_1 = \frac{\varphi_X\psi r}{2}$, $c_2 = \frac{\varphi_Y\psi r}{2}$, and $C = 2r^2(1-\psi) + 2(1-r)^2$.

Then, we have

$$\sum_{t=0}^{k} \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(t) - \bar{\boldsymbol{X}}(t)\|^2 + n\|\nabla F(\bar{X}(k))\|^2] \leq \frac{U(0)}{\theta_1}, \tag{9}$$

and

$$\mathbf{E}_{\mathcal{C}}[n(F(\bar{X}(k)) - F^\star)] < U(0), \tag{10}$$

where

$$\begin{aligned}
&\theta_1 = \min\{\theta_2, \theta_3\},\ \theta_2 = \frac{\eta}{4} - (\phi\varepsilon_1 + \varepsilon_2 + \varepsilon_3),\\
&\theta_3 = \min\{0.07(1-\sigma)\gamma, 0.44c_1(2c_1+1),\\
&\qquad 0.77c_2(2c_2+1)\},\\
&\varepsilon_1 = \frac{8L_f^2}{(1-\sigma)\gamma}\eta^2,\ \varepsilon_2 = 4(1+c_1^{-1})\eta^2,\\
&\varepsilon_3 = 5(1+c_2^{-1})\eta^2.
\end{aligned}$$

Remark 2. Theorem 1 shows that Algorithm 1 achieves a convergence rate of $\mathcal{O}(1/T)$. Specifically, (9) reveals that the term $\min_{k\leq T}\{\mathbf{E}_{\mathcal{C}}[n\|\nabla F(\bar{X}(k))\|^2 + \|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2]\}$ decays at a rate of $\mathcal{O}(1/T)$. Additionally, (10) indicates that the term $\mathbf{E}_{\mathcal{C}}[n(F(\bar{X}(k)) - F^\star)]$ is bounded.

Moreover, with Assumption 3, the following result shows that Algorithm 1 can find global optima and the convergence rate is linear.

Theorem 2. *Suppose that Assumptions 1–4 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 1 with the parameters η, γ, φ_X, and φ_Y being given in Theorem 1. Then, we have*

$$\begin{aligned}
&\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + n(F(\bar{X}(k)) - F^\star)]\\
&\leq (1-\theta_4)^k U(0),
\end{aligned} \tag{11}$$

where $\theta_4 = \min\{\theta_3, 2\nu\theta_2\}$.

Remark 3. Note that the P–Ł constant is not utilized in Algorithm 1. This is a significant property since determining the P–Ł constant can be a challenging task. It is worth noting that most existing DGT nonconvex optimization algorithms require the use of the P–Ł constant, see, e.g., Liao et al. (2023), Tang et al. (2021), Xin et al. (2021). This property of not requiring the P–Ł constant arises from the Lyapunov method, which differs from methods based on constructing systems of linear inequalities, as used in Liao et al. (2023), Tang et al. (2021), Xin et al. (2021).

3.2. Error feedback based compressed gradient tracking algorithm: Bounded relative compression error

In this section, we extend Algorithm 1 to an error feedback version for biased compressors, as shown in Algorithm 2, which is the same as the EF-C-GT proposed in Liao et al. (2022). Similar to Section 3.1, we investigate the nonconvex optimization problem without (Theorem 3) and with (Theorem 4) the P–Ł condition. In this section, we reconstruct a Lyapunov function to analyze the convergence of Algorithm 2.

Before demonstrating the convergence of Algorithm 2, we denote $\boldsymbol{E}^X = \mathrm{col}(E_1^X, \ldots, E_n^X)$, $\boldsymbol{E}^Y = \mathrm{col}(E_1^Y, \ldots, E_n^Y)$,

To analyze the convergence of Algorithm 1, we consider the following Lyapunov candidate function

$$\hat{U}(k) = \hat{V}(k) + n(F(\bar{X}(k)) - F^\star), \tag{13}$$

where $\hat{V}(k) = V(k) + \hat{\phi}(\|\boldsymbol{E}^X(k)\|^2 + \|\boldsymbol{E}^Y(k)\|^2)$, $\hat{\phi} = \frac{0.1}{C}\min\{c_1(2c_1+1), c_2(2c_2+1)\}$.

Note that Algorithm 2 introduces two error feedback variables, $\boldsymbol{E}^X$ and $\boldsymbol{E}^Y$, to rectify the bias caused by the biased compressors. Hence, the designed Lyapunov candidate function (13) incorporates two additional feedback error terms, $\|\boldsymbol{E}^X(k)\|^2$ and $\|\boldsymbol{E}^Y(k)\|^2$. Moreover, the weight parameter $\hat{\phi}$ plays a crucial role in ensuring the convergence of function (13).

Next, we investigate the convergence of Algorithm 2. Similar to Theorem 1, we first establish the following sublinear convergence result for Algorithm 2 without the P–Ł condition.

Algorithm 2

For each agent $i \in \mathcal{V}$.
Initialization:
$X_i(0) \in \mathbb{R}^d$, $Y_i(0) = \nabla F_i(X_i(0))$, $A_i(0) = B_i(0) = C_i(0) = D_i(0) = \mathbf{0}_d$, $Q_i^X(0) = \hat{Q}_i^X(0) = \mathcal{C}(X_i(0))$, and $Q_i^Y(0) = \hat{Q}_i^Y(0) = \mathcal{C}(Y_i(0))$.
Communication:
Transmit $Q_i^X(k)$, $\hat{Q}_i^X(k)$, $Q_i^Y(k)$, and $\hat{Q}_i^Y(k)$ to its out-neighbors and receive $Q_j^X(k)$, $\hat{Q}_j^X(k)$, $Q_j^Y(k)$, and $\hat{Q}_j^Y(k)$ from its in-neighbors.
Update Rule:

$$A_i(k+1) = A_i(k) + \varphi_X Q_i^X(k), \tag{12a}$$

$$B_i(k+1) = B_i(k) + \varphi_X (Q_i^X(k) - \sum_{j=1}^n W_{ij} Q_j^X(k)), \tag{12b}$$

$$C_i(k+1) = C_i(k) + \varphi_Y Q_i^Y(k), \tag{12c}$$

$$D_i(k+1) = D_i(k) + \varphi_Y (Q_i^Y(k) - \sum_{j=1}^n W_{ij} Q_j^Y(k)), \tag{12d}$$

$$X_i(k+1) = X_i(k) - \gamma [B_i(k) + \hat{Q}_i^X(k) - \sum_{j=1}^n W_{ij} \hat{Q}_j^X(k)] - \eta Y_i(k), \tag{12e}$$

$$Y_i(k+1) = Y_i(k) - \gamma [D_i(k) + \hat{Q}_i^Y(k) - \sum_{j=1}^n W_{ij} \hat{Q}_j^Y(k)] + \nabla F_i(X_i(k+1)) - \nabla F_i(X_i(k)), \tag{12f}$$

$$Q_i^X(k+1) = \mathcal{C}(X_i(k+1) - A_i(k+1)), \tag{12g}$$

$$Q_i^Y(k+1) = \mathcal{C}(Y_i(k+1) - C_i(k+1)), \tag{12h}$$

$$E_i^X(k+1) = \varsigma E_i^X(k) + X_i(k) - A_i(k) - \hat{Q}_i^X(k), \tag{12i}$$

$$E_i^Y(k+1) = \varsigma E_i^Y(k) + Y_i(k) - C_i(k) - \hat{Q}_i^Y(k), \tag{12j}$$

$$\hat{Q}_i^X(k+1) = \mathcal{C}(\varsigma E_i^X(k+1) + X_i(k+1) - A_i(k+1)), \tag{14k}$$

$$\hat{Q}_i^Y(k+1) = \mathcal{C}(\varsigma E_i^Y(k+1) + Y_i(k+1) - C_i(k+1)), \tag{14l}$$

where γ, η, φ_X, φ_Y, and ς are positive parameters.

Theorem 3. *Suppose that Assumption 1, 2, and 4 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 2 with parameters η, φ_X, and φ_Y being chosen in Theorem 1, and γ, ς, satisfying*

$$\gamma \in (0,\ \min\{\frac{c_1(1-\sigma)}{160C},\ \frac{c_1}{16\sqrt{C}},\ \frac{c_1}{20L_f\sqrt{C(1+c_2^{-1})}},\ \frac{c_2 L_f^2}{4C},\ \frac{c_2}{20\sqrt{C}},\ \frac{1}{4(1+c_1^{-1})+5(1+c_2^{-1})L_f^2},\ \frac{(1-\sigma)\hat{\phi}}{4(16(1+4\phi L_f^2)+8(1-\sigma))},\ \frac{1}{5(1+c_2^{-1})},\ \frac{(1-\sigma)\hat{\phi}}{32(2\phi+(1-\sigma))},\ \Pi\}),$$

$$\varsigma \in (0,\ \min\{\frac{1}{2\sqrt{C}},\ \frac{1}{\sqrt{2C+1}}\}). \tag{14}$$

Then, we have

$$\sum_{t=0}^k \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(t) - \bar{\boldsymbol{X}}(t)\|^2 + n\|\nabla F(\bar{X}(k))\|^2] \le \frac{\hat{U}(0)}{\hat{\theta}_1}, \tag{15}$$

and

$$\mathbf{E}_{\mathcal{C}}[n(F(\bar{X}(k)) - F^\star)] \le \mathbf{E}_{\mathcal{C}}[\hat{U}(k)] < \hat{U}(0), \tag{16}$$

where $\hat{\theta}_1 = \min\{\theta_2, \hat{\theta}_2\}$, $\hat{\theta}_2 = \min\{0.07(1-\sigma)\gamma,\ 0.24c_1(2c_1+1),\ 0.57c_2(2c_2+1),\ 0.25\}$.

Similar to Theorem 2, we then have the following linear convergence result for Algorithm 2 with Assumption 3.

Theorem 4. *Suppose that Assumptions 1–4 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 2 with parameters η, γ, ς, φ_X, and φ_Y being given in Theorem 3. Then, we have*

$$\mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + n(F(\bar{X}(k)) - F^\star)] \le (1-\hat{\theta}_3)^k U(0), \tag{17}$$

where $\hat{\theta}_3 = \min\{\hat{\theta}_2, 2\nu\theta_2\}$.

4. Compressed distributed nonconvex algorithm: Bounded absolute compression error

In this section, we propose a compressed DGT algorithm (Algorithm 3) that is designed for compressors with bounded absolute compression error, which is similar to the RCPP algorithm proposed in Liao et al. (2023). In Liao et al. (2023), the authors constructed systems of linear inequalities to analyze the convergence of RCPP algorithm using more general compressors, which allow both locally and globally absolute compression errors, under the P–Ł condition for directed graphs. However, Liao et al. (2023) requires prior knowledge of the P–Ł constant to design the algorithm parameters. In Section 4.1, we employ the Lyapunov method to demonstrate that the proposed Algorithm 3 exhibits sublinear convergence in the general nonconvex case (Theorem 5) and linear convergence when the P–Ł condition is satisfied (Theorem 6). It is important to note that the P–Ł constant is not used in designing the algorithm's parameters. In Section 4.2, we focus on compressors with locally bounded compression error and establish a linear convergence result for Algorithm 3 with the P–Ł condition (Theorem 7).

4.1. Compressed gradient tracking algorithm: Globally bounded absolute compression error

In this section, we introduce a compression operator with globally bounded absolute compression error.

Assumption 5 (*Khirirat et al., 2020; Yi et al., 2023*)**.** The compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$, adheres to the condition:

$$\mathbf{E}_{\mathcal{C}}[\|\mathcal{C}(X) - X\|_p^2] \le C,\ \forall X \in \mathbb{R}^d, \tag{19}$$

for $p \in \mathbb{Z}^+$ and constant $C \ge 0$.

Assumption 5 mainly includes deterministic quantization and unbiased random quantization. It is a commonly used compressor in the literature, as seen in Khirirat et al. (2020), Yi et al. (2023).

To analyze the convergence of Algorithm 3 using compressors that have globally bounded absolute compression error, we consider the following Lyapunov candidate function:

$$\breve{U}(k) = \breve{V}(k) + n(F(\bar{X}(k)) - F^\star), \tag{20}$$

where $\breve{V}(k) = \|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + \phi\|\boldsymbol{Y}(k) - \bar{\boldsymbol{Y}}(k)\|^2$.

From (18g) and (18h), it can be found that the compression errors are bounded by the scaling function $s(k)$. Consequently, the reconstructed Lyapunov candidate function only comprises the consensus error term, gradient tracking error term, and the optimal error term.

Algorithm 3

For each agent $i \in \mathcal{V}$.
Initialization:
$X_i(0) \in \mathbb{R}^d$, $Y_i(0) = \nabla F_i(X_i(0))$, $\hat{X}_i(-1) = \hat{Y}_i(-1) = V_i(-1) = Z_i(-1) = \mathbf{0}_d$, $Q_i^X(0) = \mathcal{C}(X_i(0)/s(0))$, $Q_i^Y(0) = \mathcal{C}(Y_i(0)/s(0))$, and $s(0) > 0$.
Communication:
Transmit $Q_i^X(k)$ and $Q_i^Y(k)$ to its out-neighbors and receive $Q_j^X(k)$ and $Q_j^Y(k)$ from its in-neighbors.
Update Rule:

$$\hat{X}_i(k) = \hat{X}_i(k-1) + s(k)Q_i^X(k), \tag{18a}$$

$$V_i(k) = V_i(k-1) + s(k)Q_i^X(k) - s(k)\sum_{j=1}^n W_{ij}Q_j^X(k), \tag{18b}$$

$$\hat{Y}_i(k) = \hat{Y}_i(k-1) + s(k)Q_i^Y(k), \tag{18c}$$

$$Z_i(k) = Z_i(k-1) + s(k)Q_i^Y(k) - s(k)\sum_{j=1}^n W_{ij}Q_j^Y(k), \tag{18d}$$

$$X_i(k+1) = X_i(k) - \gamma(\hat{X}_i(k) - V_i(k)) - \eta Y_i(k), \tag{18e}$$

$$\begin{aligned} Y_i(k+1) = {} & Y_i(k) - \gamma(\hat{Y}_i(k) - Z_i(k)) \\ & + \nabla F_i(X_i(k+1)) - \nabla F_i(X_i(k)), \end{aligned} \tag{18f}$$

$$Q_i^X(k+1) = \mathcal{C}((X_i(k+1) - \hat{X}_i(k))/s(k+1)), \tag{18g}$$

$$Q_i^Y(k+1) = \mathcal{C}((Y_i(k+1) - \hat{Y}_i(k))/s(k+1)), \tag{18h}$$

where γ, η, and μ are positive parameters, $s(k) = s(0)\mu^k > 0$ is a decreasing scaling function, and $\mu \in (0, 1)$.

Next, we investigate the convergence of Algorithm 3.

Theorem 5. *Suppose that Assumption 1, 2, and 5 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 3 with $s(0) > 0$, and μ being an arbitrary constant in $(0, 1)$, parameters η and γ being chosen in Theorem 1. Then, we have*

$$\begin{aligned} & \sum_{t=0}^k \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(t) - \bar{\boldsymbol{X}}(t)\|^2 + n\|\nabla F(\bar{X}(k))\|^2] \\ & \le \frac{\breve{U}(0) + \frac{\breve{\theta}_2}{1-\mu^2}}{\breve{\theta}_1}, \end{aligned} \tag{21}$$

and

$$\begin{aligned} & \mathbf{E}_{\mathcal{C}}[n(F(\bar{X}(k)) - F^\star)] \le \mathbf{E}_{\mathcal{C}}[\breve{U}(k)] \\ & < \breve{U}(0) + \frac{\breve{\theta}_2}{1-\mu^2}, \end{aligned} \tag{22}$$

where $\breve{\theta}_1 = \min\{\breve{\theta}_3, \breve{\theta}_4\}$, $\breve{\theta}_2 = \frac{16\gamma C n\tilde{d}^2 s^2(0)(1+2L_f^2)}{1-\sigma}$, $\breve{\theta}_3 = 0.59(1-\sigma)\gamma$, $\breve{\theta}_4 = \frac{\eta}{4} - \phi\varepsilon_1$, $\tilde{d} = 1$ *for* $p \in [1, 2]$, *and* $\tilde{d} = d^{\frac{1}{2}-\frac{1}{p}}$ *for* $p > 2$, *with* $p \in \mathbb{Z}^+$ *representing the constant of the p-norm.*

Similar to Theorem 2, we then have the following linear convergence result for Algorithm 3 with Assumption 3.

Theorem 6. *Suppose that Assumptions 1–3 and 5 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 3 with $s(0) > 0$, and μ being an arbitrary constant in $(0, 1)$, parameters η and γ being chosen in Theorem 1. Then, we have*

$$\begin{aligned} & \mathbf{E}_{\mathcal{C}}[\|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + n(F(\bar{X}(k)) - F^\star)] \\ & < (1 - \breve{\theta}_5)^k \breve{\theta}_6, \end{aligned} \tag{23}$$

where $\breve{\theta}_5 = \min\{\breve{\theta}_7, 1-\mu^2\}$, *and*

$$\breve{\theta}_6 = \breve{U}(0) + \breve{\theta}_8 s^2(0) \begin{cases} \frac{1}{(1-\breve{\theta}_9)\mu^2}, & \text{if } 1-\breve{\theta}_7 < \mu^2, \\ \frac{1}{(1-\breve{\theta}_{10})(1-\breve{\theta}_7)}, & \text{if } 1-\breve{\theta}_7 > \mu^2, \\ \frac{1}{(1-\breve{\theta}_{11})\varpi}, & \text{if } 1-\breve{\theta}_7 = \mu^2, \end{cases}$$

$$\breve{\theta}_7 = \min\{\breve{\theta}_1, 2\nu\theta_2\},\ \breve{\theta}_8 = \frac{16n\tilde{d}^2\gamma C}{1-\sigma}(1 + 2L_f^2),$$

$$\breve{\theta}_9 = \frac{1-\breve{\theta}_7}{\mu^2},\ \breve{\theta}_{10} = \frac{\mu^2}{1-\breve{\theta}_7},\ \breve{\theta}_{11} = \frac{\mu^2}{\varpi},\ \varpi \in (\mu^2, 1).$$

4.2. Compressed gradient tracking algorithm: Locally bounded absolute compression error

In this section, we introduce a compression operator with locally bounded absolute compression error.

Assumption 6 (Yi et al., 2023; Zhang et al., 2023)**.** The compression operator $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$, adheres to the condition:

$$\begin{aligned} & \|\mathcal{C}(X) - X\|_p \le (1-\varphi), \\ & \forall X \in \{X \in \mathbb{R}^d : \|X\|_p \le 1\}, \end{aligned} \tag{24}$$

for $p \in \mathbb{Z}^+$ and constant $\varphi \in (0, 1]$.

Assumption 6 mainly includes standard quantization with dynamic and fixed quantization levels, which are commonly used compression techniques in the literature, see Yi et al. (2023), Zhang et al. (2023).

To analyze the convergence of Algorithm 3 using compressors that have locally bounded absolute compression error, we consider the following Lyapunov candidate function:

$$\tilde{U}(k) = \breve{V}(k) + \tilde{\phi} n(F(\bar{X}(k)) - F^\star), \tag{25}$$

where $\tilde{\phi} = \frac{0.4\gamma(1-\sigma)}{\eta L_f^2}$.

Next, we investigate the convergence of Algorithm 3.

Theorem 7. *Suppose that Assumptions 1–3 and 6 hold. Let each agent $i \in \mathcal{V}$ run Algorithm 3 with the following parameters:*

$$\begin{aligned} \eta \in & (0, \min\{\frac{(1-\sigma)^2\gamma}{40L_f},\ \frac{\varphi+\varphi^2-\varphi^3}{2\tilde{\xi}_1}, \\ & \frac{\varphi+\varphi^2-\varphi^3}{2\tilde{\xi}_2}, 1\}), \\ \gamma \in & (0, \min\{\sqrt{\frac{\varphi+\varphi^2-\varphi^3}{2\tilde{\xi}_3}},\ \sqrt{\frac{\varphi+\varphi^2-\varphi^3}{2\tilde{\xi}_4}}, \\ & \frac{2L_f^2}{(1-\sigma)\nu}\}), \\ s(0) \ge & \max\{\sqrt{\frac{\tilde{U}(0)}{\tilde{\xi}_5}}, \max_{i\in\mathcal{V}}\|X_i(0)\|, \max_{i\in\mathcal{V}}\|Y_i(0)\|\}, \\ \mu \in & [\max\{\sqrt{\tilde{\theta}_1},\ \sqrt{\tilde{\xi}_6},\ \sqrt{\tilde{\xi}_7}\},\ 1), \end{aligned} \tag{26}$$

where

$$\tilde{\theta}_1 = 1 - \tilde{\theta}_3 + \frac{\tilde{\theta}_2}{\tilde{\xi}_5}\gamma,\ \tilde{\theta}_2 = 2(1+2L_f^2)\frac{8n\tilde{d}^2(1-\varphi)^2}{1-\sigma},$$

$$\tilde{\theta}_3 = \tilde{\theta}_4\gamma,\ \tilde{\theta}_4 = \min\{0.59(1-\sigma),\ \frac{48\nu\phi}{1-\sigma}\},$$

$$\tilde{\xi}_1 = 4\hat{d}^2(5+4L_f^2)(1+\varphi^{-1})\tilde{\xi}_5,$$

$$\tilde{\xi}_2 = 10L_f^2(3+2L_f^2)(1+\varphi^{-1})\tilde{\xi}_5,$$

$$\tilde{\xi}_3 = \tilde{\xi}_8(1-\varphi)^2 + 32\hat{d}^2(1+\varphi^{-1})\tilde{\xi}_5,$$
$$\tilde{\xi}_4 = \tilde{\xi}_9(1+L_f^2)(1-\varphi)^2 + 40\hat{d}^2(1+L_f^2)(1+\varphi^{-1})\tilde{\xi}_5,$$
$$\tilde{\xi}_5 > \frac{\tilde{\theta}_2}{\tilde{\theta}_4},\ \tilde{\xi}_6 = (1-(\varphi+\varphi^2-\varphi^3)+\eta\tilde{\xi}_1+\gamma^2\tilde{\xi}_3),$$
$$\tilde{\xi}_7 = (1-(\varphi+\varphi^2-\varphi^3)+\eta\tilde{\xi}_2+\gamma^2\tilde{\xi}_4),$$
$$\tilde{\xi}_8 = 16n\hat{d}^2\breve{d}^2(1+\varphi^{-1}),\ \tilde{\xi}_9 = 20n\hat{d}^2\breve{d}^2(1+\varphi^{-1}),$$

and $\hat{d} = d^{\frac{1}{2}-\frac{1}{p}}$ *for* $p \in [1,2]$, $\hat{d} = 1$ *for* $p > 2$, *with* $p \in \mathbb{Z}^+$ *representing the constant of the p-norm.*

Then, we have

$$\|\boldsymbol{X}(k) - \bar{\boldsymbol{X}}(k)\|^2 + n(F(\bar{X}(k)) - F^\star) \le \tilde{\xi}_5 s^2(k). \tag{27}$$

Remark 4. The standard uniform quantizer, as used in Ma et al. (2021), Xiong et al. (2022) for the strongly convex case, and in Xu et al. (2022) for the nonconvex case under the P–Ł condition, is a widely used method for reducing communication overhead in distributed optimization. Note that Assumption 6 serves as a more general compressor. In other words, Theorem 7 demonstrates that the proposed algorithm achieves linear convergence with a broader range of compressors and only requires the global cost function to satisfy the P–Ł condition.

5. Conclusions

In this paper, we introduced three classes of compressors to reduce the communication overhead. By integrating them with DGT algorithm, we then proposed three distributed algorithms with compressed communication for distributed nonconvex optimization. For the case where local cost functions are smooth, we designed several Lyapunov functions to demonstrate that the proposed algorithms sublinearly converge to a stationary point. Moreover, when the global cost function satisfies the P–Ł condition, we demonstrated that the proposed algorithm converges linearly to a global optimal point. One future direction is to investigate general unbalanced directed graphs.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3305904, the National Natural Science Foundation of China under Grants 62133003, 61991403, 61991400, 62325304 & U22B2046, the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant No. BK20233002, the Knut and Alice Wallenberg Foundation, Sweden, the Fundamental Research Funds for the Central Universities, China 08002150267, the Shanghai Municipal Science and Technology Major Project 2021SHZDZX0100, and the Swedish Foundation for Strategic Research.

References

Alghunaim, S. A., & Sayed, A. H. (2020). Linear convergence of primal–dual gradient methods and their performance in distributed optimization. *Automatica*, *117*, Article 109003.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 1707–1718.

Cao, X., Başar, T., Diggavi, S., Eldar, Y. C., Letaief, K. B., Vincent Poor, H., & Zhang, J. (2023). Communication-efficient distributed learning: An overview. *IEEE Journal on Selected Areas in Communications*, *41*(4), 851–873.

Carnevale, G., & Notarstefano, G. (2022). Nonconvex distributed optimization via Lasalle and singular perturbations. *IEEE Control Systems Letters*, *7*, 301–306.

Di Lorenzo, P., & Scutari, G. (2016). Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, *2*(2), 120–136.

Doan, T. T., Maguluri, S. T., & Romberg, J. (2020). Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach. *IEEE Transactions on Automatic Control*, *66*(10), 4469–4484.

Du, W., Yao, L., Wu, D., Li, X., Liu, G., & Yang, T. (2018). Accelerated distributed energy management for microgrids. In *2018 IEEE power & energy society general meeting*.

Forero, P. A., Cano, A., & Giannakis, G. B. (2011). Distributed clustering using wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, *5*(4), 707–724.

Guo, J., Hug, G., & Tonguz, O. K. (2016). A case for nonconvex distributed optimization in large-scale power systems. *IEEE Transactions on Power Systems*, *32*(5), 3842–3851.

Hong, M., Hajinezhad, D., & Zhao, M.-M. (2017). Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International conference on machine learning* (pp. 1529–1538).

Horváth, S., Kovalev, D., Mishchenko, K., Richtárik, P., & Stich, S. (2023). Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods & Software*, *38*(1), 91–106.

Khirirat, S., Magnússon, S., & Johansson, M. (2020). Compressed gradient methods with Hessian-aided error compensation. *IEEE Transactions on Signal Processing*, *69*, 998–1011.

Kia, S. S., Cortés, J., & Martínez, S. (2015). Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, *55*, 254–264.

Koloskova, A., Lin, T., & Stich, S. U. (2021). An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 11422–11435.

Liao, Y., Li, Z., Huang, K., & Pu, S. (2022). A compressed gradient tracking method for decentralized optimization with linear convergence. *IEEE Transactions on Automatic Control*, *67*(10), 5622–5629.

Liao, Y., Li, Z., & Pu, S. (2023). A linearly convergent robust compressed push-pull method for decentralized optimization. arXiv preprint arXiv:2303.07091.

Lin, T., Karimireddy, S. P., Stich, S., & Jaggi, M. (2021). Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *International conference on machine learning* (pp. 6654–6665).

Ma, X., Yi, P., & Chen, J. (2021). Distributed gradient tracking methods with finite data rates. *Journal of Systems Science and Complexity*, *34*(5), 1927–1952.

Nedić, A., & Liu, J. (2018). Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, *1*, 77–103.

Nedić, A., Olshevsky, A., & Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, *27*(4), 2597–2633.

Nedić, A., & Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, *54*(1), 48–61. http://dx.doi.org/10.1109/TAC.2008.2009515.

Notarnicola, I., Bin, M., Marconi, L., & Notarstefano, G. (2023). The gradient tracking is a distributed integral action. *IEEE Transactions on Automatic Control*, *68*(12), 7911–7918.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., & Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International conference on machine learning* (pp. 2681–2690).

Qu, G., & Li, N. (2017). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, *5*(3), 1245–1260.

Reisizadeh, A., Taheri, H., Mokhtari, A., Hassani, H., & Pedarsani, R. (2019). Robust and communication-efficient collaborative learning. *Advances in Neural Information Processing Systems*, 8388–8399.

Sayed, A. H. (2014). Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, *7*, 311–801.

Shi, W., Ling, Q., Wu, G., & Yin, W. (2015). EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, *25*(2), 944–966.

Shi, Y., Yang, K., Jiang, T., Zhang, J., & Letaief, K. B. (2020). Communication-efficient edge AI: Algorithms and systems. *IEEE Communications Surveys & Tutorials*, *22*(4), 2167–2191.

Taheri, H., Mokhtari, A., Hassani, H., & Pedarsani, R. (2020). Quantized decentralized stochastic learning over directed graphs. In *International conference on machine learning* (pp. 9324–9333).

Tang, Y., Zhang, J., & Li, N. (2021). Distributed zero-order algorithms for nonconvex multi-agent optimization. *IEEE Transactions on Control of Network Systems*, *8*(1), 269–281.

Varagnolo, D., Zanella, F., Cenedese, A., Pillonetto, G., & Schenato, L. (2015). Newton-Raphson consensus for distributed convex optimization. *IEEE Transactions on Automatic Control*, *61*(4), 994–1009.

Wang, J., & Elia, N. (2010). Control approach to distributed optimization. In *Annual allerton conference on communication, control, and computing* (pp. 557–561).

Xin, R., Khan, U. A., & Kar, S. (2021). An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, *69*, 1842–1858.

Xiong, Y., Wu, L., You, K., & Xie, L. (2022). Quantized distributed gradient tracking algorithm with linear convergence in directed networks. *IEEE Transactions on Automatic Control*, *68*(9), 5638–5645.

Xu, L., Yi, X., Sun, J., Shi, Y., Johansson, K. H., & Yang, T. (2022). Quantized distributed nonconvex optimization algorithms with linear convergence. arXiv preprint arXiv:2207.08106.

Xu, L., Yi, X., Wen, G., Shi, Y., Chai, T., Johansson, K. H., & Yang, T. (2023). Compressed gradient tracking algorithms for distributed nonconvex optimization. arXiv preprint arXiv:2310.18871.

Xu, J., Zhu, S., Soh, Y. C., & Xie, L. (2015). Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE conference on decision and control* (pp. 2055–2060).

Xu, J., Zhu, S., Soh, Y. C., & Xie, L. (2017). Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, *63*(2), 434–448.

Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., & Johansson, K. H. (2019). A survey of distributed optimization. *Annual Reviews in Control*, *47*, 278–305.

Yi, P., & Hong, Y. (2014). Quantized subgradient algorithm and data-rate analysis for distributed optimization. *IEEE Transactions on Control of Network Systems*, *1*(4), 380–392. http://dx.doi.org/10.1109/TCNS.2014.2357513.

Yi, X., Zhang, S., Yang, T., Chai, T., & Johansson, K. H. (2021). Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization. *IEEE Transactions on Automatic Control*, *67*(8), 4194–4201.

Yi, X., Zhang, S., Yang, T., Chai, T., & Johansson, K. H. (2023). Communication compression for distributed nonconvex optimization. *IEEE Transactions on Automatic Control*, *68*(9), 5477–5492.

Yuan, K., Chen, Y., Huang, X., Zhang, Y., Pan, P., Xu, Y., & Yin, W. (2021). DecentLaM: Decentralized momentum SGD for large-batch deep training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3029–3039).

Yuan, D., Xu, S., Zhao, H., & Rong, L. (2012). Distributed dual averaging method for multi-agent optimization with quantized communication. *Systems & Control Letters*, *61*(11), 1053–1061.

Zhang, C., & Wang, Y. (2019). Sensor network event localization via nonconvex nonsmooth ADMM and augmented lagrangian methods. *IEEE Transactions on Control of Network Systems*, *6*(4), 1473–1485.

Zhang, J., You, K., & Başar, T. (2018). Distributed discrete-time optimization in multiagent networks using only sign of relative state. *IEEE Transactions on Automatic Control*, *64*(6), 2352–2367.

Zhang, J., You, K., & Xie, L. (2023). Innovation compression for communication-efficient distributed optimization with linear convergence. *IEEE Transactions on Automatic Control*, *68*(11), 6899–6906.

Zhao, H., Li, B., Li, Z., Richtárik, P., & Chi, Y. (2022). BEER: Fast $\mathcal{O}(1/T)$ rate for decentralized nonconvex optimization with communication compression. *Advances in Neural Information Processing Systems*, 31653–31667.

Lei Xu received the B.S. and M.S. degrees in control theory and engineering from LiaoNing Petrochemical University, Fushun, China, in 2017 and 2020, respectively. Since 2020, he has been working toward the Ph.D. degree at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. Since 2023, he has been working toward the Ph.D. degree at the Department of Mechanical Engineering, University of Victoria, Victoria, Canada. His research focuses on distributed optimization, networked control systems, event-triggered control, and Markovian jump systems.

Xinlei Yi received the B.S. and M.S. degrees in mathematics from 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 KTH Royal Institute of Technology, Stockholm, Sweden, in 2020. He was a Postdoc with KTH Royal Institute of Technology from 2020 to 2022 and with the Lab for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA, from 2022 to 2024. He is a tenure-track professor of Shanghai Institute of Intelligent Science and Technology, Tongji University. Dr. Yi was selected as one of the four finalists for the 2021 European Systems & Control Ph.D. Thesis Award. His current research interests include distributed online and optimization, online optimization, meta-learning and graph neural networks.

Guanghui Wen received the Ph.D. degree in mechanical systems and control from Peking University, Beijing, China, in 2012. He is currently an Endowed Chair Professor at Department of Systems Science, Southeast University, Nanjing, China. His current research interests include coordination control of autonomous intelligent systems, analysis and synthesis of complex networks, cyber–physical systems, resilient control, and distributed reinforcement learning.

Prof. Wen was the recipient of the National Science Fund for Distinguished Young Scholars, Australian Research Council Discovery Early Career Researcher Award, and Asia Pacific Neural Network Society Young Researcher Award. He is a reviewer for American Mathematical Review and is an active reviewer for many journals. He currently serves as an Associate Editor of the IEEE Transactions on Control of Network Systems, the IEEE Transactions on Industrial Informatics, the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Journal of Emerging and Selected Topics in Industrial Electronics, the IEEE Transactions on Systems, Man and Cybernetics: Systems, the IEEE Open Journal of the Industrial Electronics Society, and the Asian Journal of Control. Prof. Wen has been named a Highly Cited Researcher by Clarivate Analytics since 2018. He is an IET Fellow.

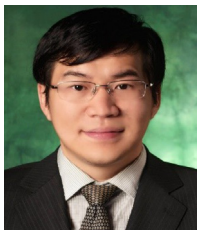
Yang Shi received his B.Sc. and Ph.D. degrees in mechanical engineering and automatic control from Northwestern Polytechnical University, Xi'an, China, in 1994 and 1998, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2005. He was a Research Associate in the Department of Automation, Tsinghua University, China, during 1998–2000. From 2005 to 2009, he was an Assistant Professor and Associate Professor in the Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada. In 2009, he joined the University of Victoria, and now he is a Professor in the Department of Mechanical Engineering, University of Victoria, Victoria, BC, Canada. His current research interests include networked and distributed systems, model predictive control (MPC), cyber–physical systems (CPS), robotics and mechatronics, navigation and control of autonomous systems (AUV and UAV), and energy system applications.

On teaching and mentorship, Dr. Shi received the University of Saskatchewan Student Union Teaching Excellence Award in 2007, and the Faculty of Engineering Teaching Excellence Award in 2012 at the University of Victoria (UVic), and the 2023 REACH Award for Excellence in Graduate Student Supervision and Mentorship. On research, he is the recipient of the JSPS Invitation Fellowship (short-term) in 2013, the UVic Craigdarroch Silver Medal for Excellence in Research in 2015, the 2017 IEEE Transactions on Fuzzy Systems Outstanding Paper Award, the Humboldt Research Fellowship for Experienced Researchers in 2018; CSME Mechatronics Medal (2023); IEEE Dr.-Ing. Eugene Mittelmann Achievement Award (2023). He is IFAC Council Member; VP on Conference Activities of IEEE IES and the Chair of IEEE IES Technical Committee on Industrial Cyber-Physical Systems. Currently, he is Editor-in-Chief of IEEE Transactions on Industrial Electronics, and Editor-in-Chief of IEEE Canadian Journal of Electrical and Computer Engineering; he also serves as Associate Editor for Automatica, IEEE Transactions on Automatic Control, Annual Review in Controls, etc. He is a Distinguished Lecturer of IES.

He is a Fellow of IEEE, ASME, CSME, Engineering Institute of Canada (EIC), Canadian Academy of Engineering (CAE), and a registered Professional Engineer in British Columbia, Canada.

Karl H. Johansson is Swedish Research Council Distinguished Professor in Electrical Engineering and Computer Science at KTH Royal Institute of Technology in Sweden and Founding Director of Digital Futures. He earned his M.Sc. degree in Electrical Engineering and Ph.D. in Automatic Control from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU and other prestigious institutions. His research interests focus on networked control systems and cyber–physical systems with applications in transportation, energy, and automation networks. For his scientific contributions, he has received numerous best paper awards and various other distinctions from IEEE, IFAC, and other organizations. He has been awarded Distinguished Professor by the Swedish Research Council, Wallenberg Scholar by the Knut and Alice Wallenberg Foundation, Future Research Leader by the Swedish Foundation for Strategic Research. He has also received the triennial IFAC Young Author Prize, IEEE CSS Distinguished Lecturer, IFAC Outstanding Service Award, and IEEE CSS Hendrik W. Bode Lecture Prize. His extensive service to the academic community includes being President of the European Control Association, IEEE CSS Vice President Diversity, Outreach & Development, and

Member of IEEE CSS Board of Governors and IFAC Council. He has served on the editorial boards of Automatica, IEEE TAC, IEEE TCNS and many other journals. He has also been a member of the Swedish Scientific Council for Natural Sciences and Engineering Sciences. He is Fellow of both the IEEE and the Royal Swedish Academy of Engineering Sciences.



Tao Yang is a Professor at the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University. He was an Assistant Professor at the Department of Electrical Engineering, University of North Texas, Denton, USA, from 2016-2019. He received the Ph.D. degree in electrical engineering from Washington State University in 2012. Between August 2012 and August 2014, he was an ACCESS postdoctoral researcher with the ACCESS Linnaeus Centre, 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. 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 multi-agent systems. He is an Associate Editor for IEEE Transactions on Control of Network Systems, IEEE Transactions on Control Systems Technology, and IEEE Transactions on Neural Networks and Learning System. He received Ralph E. Powe Junior Faculty Enhancement Award and Best Student Paper award (as an advisor) of several international Conference.