Convergence of an accelerated distributed optimisation algorithm over time-varying directed networks

Jinhui Hu1 | Yu Yan1 | Huaqing Li1 | Zheng Wang1 | Dawen Xia2 | Jing Guo1

1 Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing 400715, PR China
2 College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, PR China

Correspondence
Huaqing Li, Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing, 400715 PR China. Email: huaqingli@swu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61773321; Central Universities in China, Grant/Award Number: XDJK2019AC001; Venture and Innovation Support Program for Chongqing Overseas Returns, Grant/Award Number: cx2019905; China Scholarship Council, Grant/Award Number: 201808525063; Science and Technology Talents Fund for Excellent Young of Guizhou, Grant/Award Number: QKHPTRC20195669; Guizhou Science and Technology Department, Grant/Award Number: QKHJIC20181083; Science and Technology Top-notch Talents Support Project of Colleges and Universities in Guizhou, Grant/Award Number: QJHKY2016665; High-level Innovative Talents Project of Guizhou, Grant/Award Numbers: QRLF201621, 61762020

Abstract
In this article, studying distributed optimisation over time-varying directed networks where a group of agents aims at cooperatively minimising a sum of local objective functions is focused on. Each agent uses only local computation and communication in the overall process without leaking their private information. Via incorporating both a distributed heavy-ball method and a distributed Nesterov method, a double accelerated distributed algorithm leveraging a gradient-tracking technique and using uncoordinated step-sizes, is developed. By employing both row- and column-stochastic weight matrices, the proposed algorithm can bypass the implementation of doubly stochastic weight matrices and avoid eigenvector estimation existing in some algorithms using only row- or column-stochastic weight matrices. Under the assumptions that the agents’ local objective functions are smooth and strongly convex, and the aggregated directed networks of every finite consecutive directed network are strongly connected, the proposed algorithm is proved to converge linearly to the global optimal solution when the largest step-size is positive and sufficiently small, and the largest momentum parameter is non-negative. The proposed algorithm is also applied to fixed directed networks which are considered as a special case of time-varying directed networks. Simulation results further verify the effectiveness of the proposed algorithm and correctness of the theoretical findings.

1 | INTRODUCTION

In this paper, we consider a network of $m$ agents to collaboratively resolve the following optimisation problem

$$\min_{\tilde{x} \in \mathbb{R}^n} \tilde{f}(\tilde{x}) = \frac{1}{m} \sum_{i=1}^{m} \tilde{f}_i(\tilde{x}),$$

(1)

where each local objective function, $\tilde{f}_i: \mathbb{R}^n \rightarrow \mathbb{R}$, is privately known by agent $i$. Via exchanging information with nearby agents, the mutual goal of the agents is to seek the optimal solution, $\tilde{x}^* \in \mathbb{R}^n$, to problem (1).

Nowadays, optimisation problems with form (1) have attracted considerable attention in many fields, such as machine learning [1, 2], signal processing [3, 4], coordinated control [5], sensor networks [6, 7], and EDP [8, 9]. Early works tackling this kind of problems include Distributed Gradient Descent (DGD) [10, 11] over undirected networks, which attain a relatively slow convergence rate due to the diminishing step-sizes. The convergence rates are $O(\log k/\sqrt{k})$ for general convex
objective functions and $\mathcal{O}(\log k/k)$ for strongly convex objective functions, where $k$ is the iteration number. Then, Jadabaie et al. [5] leverage a constant step-size to achieve faster convergence rate but inexact convergence. In addition, some algorithms based on Lagrangian dual [12–14] also achieve faster convergence at the cost of higher computational burden. In order to improve convergence and attain computational simplicity, Jakovetic et al. [15] utilise a Nesterov method to reach convergence rate at $\mathcal{O}(1/k)$ for general convex objective functions and linear convergence rate for strongly convex objective functions. The literature [16–18] construct only doubly stochastic weight matrices for distributed optimisation over undirected networks. However, in practice, the communications in networks may be directed, which means that the weight matrices cannot be doubly stochastic due to the imbalanced transmission of the information among agents.

The existing methods tackling distributed optimisation problems over directed networks are developed by the methods proposed in [16, 19, 20] that realise the imbalanced information transmission among agents by adopting average-consensus methods. Nevertheless, these algorithms, limited by diminishing step-sizes, slowly converge to the global optimal solution at a sublinear rate. In order that improve the convergence rate, DEXTRA [21] combines a push-sum technique with EXTRA to attain a linear convergence rate over directed networks when the step-size is chosen from a small interval. An elegant work Push-DIGing [22] extends ADD-OPT [23] to time-varying networks and achieves linear convergence rate by applying the small gain theorem [24]. For eliminating the deficiency of the use of column-stochastic weight matrix, FROST [25] utilises only row-stochastic weights, the implementation of which does not need the agent knowing its out-degree. More recently, Xin et al. [26] provide a novel method, $AB$, which simultaneously employs row- and column-stochastic weights to remove limitations caused by the doubly stochastic weights and further simplify the computation and communication process. Then, TV-$AB$ [27] extends the $AB$ to time-varying directed networks. For achieving accelerated linear convergence, the distributed heavy-ball method $ABm$ [28] adds a momentum term into the $AB$ algorithm. Note that the $ABm$ algorithm is based on a gradient-tracking method first used in [29], which is built on dynamic consensus and renders the agents to approach gradients of the global objective function step-wisely. Furthermore, $ABm$ algorithm leverages the distributed heavy-ball method to effectively accelerate the convergence. Qu et al. [30] propose a distributed Nesterov method over fixed undirected networks.

Motivated by the insightful work [27], we aim to incorporate both the distributed heavy-ball method [28] and the distributed Nesterov method [30, 31] into [27] to technically achieve double acceleration. Moreover, the researchers in [27] do not provide an explicit bound on the constant step-size limited by the analysis technique while the explicit bounds on the step-size have been analysed by many works [3, 23, 28, 32–34], even by the time-varying work Push-DIGing [22]. Therefore, this paper aims at closing this gap. The main contributions of this paper are summarised as follows:

i. We incorporate both the distributed heavy-ball method [28] and the distributed Nesterov method [31] into the proposed algorithm. Both the distributed heavy-ball method and distributed Nesterov method are testified to be able to accelerate the convergence. Thus, the proposed algorithm realises double acceleration in comparison with some existing algorithms [4, 12–16, 19–23, 25–27, 29].

ii. In contrast with the time-varying algorithms Push-DIGing [22] and TV-$AB$ [27], the proposed algorithm employs uncoordinated step-sizes, which allows each agent in the network to locally decide its own constant steps-size and thus increases the autonomy and flexibility of the agents. Based on the gradient-tracking method, the proposed algorithm is proved to converge linearly to the global optimal solution over time-varying directed networks. Furthermore, the explicit upper bounds on the largest step-size and the largest momentum parameter are derived, which are significant to practical applications.

iii. The proposed algorithm simultaneously utilises row- and column-stochastic weight matrices, which eliminates the requirement of using doubly stochastic weight matrices existing in [10, 11, 16]. Furthermore, compared with recently insightful works FROST [25] and Push-DIGing [22], although we simultaneously utilise row- and column-stochastic weight matrices, the proposed algorithm can avoid the process of eigenvector estimation at each iteration, which reduces the computational complexity and communication burden.

iv. By applying the proposed double accelerated distributed optimisation algorithm into fixed directed networks, we prove that the fixed directed networks [35] can be considered as a special case of time-varying directed networks. In the case of fixed directed networks, the proposed algorithm can still accelerate the linear convergence in [28] owing to the distributed Nesterov method, which can be observed in Section 4.3.

The remainder of this paper is organised in the following. Section 2 presents some preliminaries and algorithm development. Section 3 provides rigorous convergence analysis. Some real-world data sets are utilised in Section 4 to help conduct experiments. We draw a conclusion and state our future work in Section 5. Some crucial proofs are placed in Appendix to ensure the coherence of the whole paper.
2 | PRELIMINARIES AND ALGORITHM DEVELOPMENT

2.1 | Notations

All vectors in this paper are recognised as column vectors if no otherwise specified. The $n \times n$ identity matrix are denoted by $I_n$ and so forth. We denote by $1_m$ the $m$-dimensional vector with all-one elements. We use $x^T$ and $A^T$ to respectively denote the transpose of vector $x$ and the transpose of matrix $A$, and $A^{ij}$ denotes the $(i,j)$-th element of matrix $A$. For an arbitrary vector $x$, we denote its $i$-th element by $[x]_i$ and denote the largest element of vector $x$ by $\|x\|_{\max}$, while $\text{diag}(x)$ represents a diagonal matrix with all the element of $x$ laying on its main diagonal. For two matrices, $X$ and $Y$ with same dimensions; $X \leq Y$ means that each element in $Y - X$ is nonnegative; $X \otimes Y$ represents their Kronecker product; the spectral radius of matrix $X$ is denoted by $\rho(X)$. The notation $\| \cdot \|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. For two arbitrary vectors, $x_i, y_j \in \mathbb{R}^m$, we introduce two weighted inner products, $\langle x_i, y_j \rangle_{\pi_i} \triangleq x_i^T ((\text{diag}(\pi_j))^T \otimes I_m) y_j$ and $\langle x_i, y_j \rangle_{\pi_i} \triangleq x_i^T ((\text{diag}(\pi_j))^{-1} \otimes I_m) y_j$, which leads to two weighted Euclidean norms, $\| x \|_{\pi_i} \triangleq \|(\text{diag}(\sqrt{\pi_j}))^{-1} \otimes I_m) x \|$ and $\| x \|_{\pi_i} \triangleq \|(\text{diag}(\sqrt{\pi_j})) \otimes I_m) x \|$.

2.2 | Problem reformulation

Assume that a system of $m$ agents communicate over a time-varying directed networks, $\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k)$, where $\mathcal{V} = \{1, \ldots, m\}$ is the set of agents and $k$ is the iteration number. $\mathcal{E}_k$ is the set of directed edges at time $k$, such that if $(i, j) \in \mathcal{E}_k$, agent $i$ can transmit messages to agent $j$, that is, $i \rightarrow j$, at iteration $k$. The common target of agents is to resolve a consensus problem as follows:

$$\min_{x \in \mathbb{R}^m} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x^i)$$

s.t. $x^i = x^j, \forall i, j \in \mathcal{E}_k,$ \hspace{1cm} (2)

where $x = [x^1^T, \ldots, x^m^T]^T$ and each local objective function, $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, is only known by agent $i$. Let $x^* = 1_m \otimes x^*$ denote the global optimal solution to problem (2). Some necessary assumptions and lemmas in this paper are presented in the following.

**Assumption 1.** Each local objective function $\tilde{f}_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \ldots, m$, is strongly convex with strong convexity parameter $\mu$, and it has Lipschitz continuous gradient with Lipschitz constant $L_{\tilde{f}_i}$, where $L_{\tilde{f}_i} \geq \mu > 0$. That is, for all $i = 1, \ldots, m$, and $\tilde{x}, \tilde{y} \in \mathbb{R}^n$, it holds

$$\| \nabla \tilde{f}_i (\tilde{x}) - \nabla \tilde{f}_i (\tilde{y}) \| \leq L_{\tilde{f}_i} \| \tilde{x} - \tilde{y} \|,$$

$$\tilde{f}_i (\tilde{x}) - \tilde{f}_i (\tilde{y}) \geq \nabla \tilde{f}_i^T (\tilde{y}) (\tilde{x} - \tilde{y}) + \frac{\mu}{2} \| \tilde{x} - \tilde{y} \|^2.$$ \hspace{1cm} (4)

**Assumption 2.** For a time-varying network sequence $\{\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k) : k \geq 0\}$, there exist some positive integers $C$ such that for any $k \geq 0$, the aggregate directed network $\mathcal{G}_k \coloneqq (\mathcal{V}, \bigcup_{l=0}^k \mathcal{E}_l)$ is strongly connected.

**Assumption 3.** $\{\mathcal{G}_k = (\mathcal{V}, \mathcal{E}_k)\}$ is a sequence of time-varying directed networks, and the matrix sequences, $\{A_k\}$ and $\{B_k\}$, aligned with $(i, j) \in \mathcal{E}_k \iff A_{k}^{ij} B_{k}^{ij} \neq 0$, the following assumptions satisfy. (1) $\{A_k\}$ and $\{B_k\}$ are respectively row- and column-stochastic weight matrices, that is, $\sum_{j=1}^{n} A_{k}^{ij} = 1, \forall i \in \mathcal{V}$ and $\sum_{i=1}^{m} B_{k}^{ij} = 1, \forall j \in \mathcal{V}$; (2) $\mathcal{G}_k$ has self-loops; that is, $A_{k}^{ii} > 0$ and $B_{k}^{ii} > 0, \forall i \in \mathcal{V}$ and $\forall k \geq 0$; (3) There are scalars $0 < w_1, w_2 < 1$ such that $A_{k}^{ij} > w_1$ and $B_{k}^{ij} > w_2$, $(i, j) \in \mathcal{E}_k, \forall k \geq 0$.

**Remark 1.** Assumption 1 guarantees that the optimal solution to problem (1) exists uniquely and Assumption 2 ensures the C-bounded strong-connectivity of the communication networks. Notice that the uniform positivity bounds $w_1$ and $w_2$ in Assumption 3 are not necessarily known by any of the agents.

2.3 | The heavy-ball method

The centralised gradient descent method proposed in [36] is

$$\tilde{x}_{k+1} = \tilde{x}_k - \alpha \nabla \tilde{f}(\tilde{x}_k),$$

where $\tilde{x}_k \in \mathbb{R}^n$ and $\alpha$ serves as a positive constant step-size. This method achieves the best convergence rate at $\mathcal{O}\left((\frac{\mu}{\alpha} + 1)^{-k}\right)$, where $\mathcal{Q}$ is the condition number of global objective function $f$. Clearly, gradient descent is quite slow when $\mathcal{Q}$ is large, that is, the objective function is ill-conditioned [28]. To address this issue, Polyak in [37] proposes a centralised heavy-ball method as follows:

$$\tilde{x}_{k+1} = \tilde{x}_k - \alpha \nabla \tilde{f}(\tilde{x}_k) + \tilde{\beta} (\tilde{x}_k - \tilde{x}_{k-1}),$$

where $\tilde{\beta} (\tilde{x}_k - \tilde{x}_{k-1})$ is referred as a momentum term with the positive momentum parameter $\tilde{\beta}$, which is utilised to accelerate the convergence process. [31] achieves an accelerated rate of $\mathcal{O}\left((\frac{\mu}{\alpha} + 1)^{-k}\right)$ with specific choices of $\alpha$ and $\tilde{\beta}$.

2.4 | The Nesterov method

Based on (5), Nesterov in [38] proposes an efficient centralised Nesterov method to accelerate the convergence of (5) through an added momentum term, the details of which are as follows:

$$\tilde{y}_{k+1} = \tilde{x}_k - \alpha \nabla \tilde{f}(\tilde{x}_k),$$

$$\tilde{x}_{k+1} = \tilde{y}_{k+1} + \tilde{\beta} (\tilde{y}_{k+1} - \tilde{x}_k),$$

where $\tilde{y}_{k+1} = \tilde{x}_k - \alpha \nabla \tilde{f}(\tilde{x}_k), \tilde{x}_{k+1} = \tilde{y}_{k+1} + \tilde{\beta} (\tilde{y}_{k+1} - \tilde{x}_k).$
where $y_k \in \mathbb{R}^n$ is a variable and is corrected in the next update. Refer to [31], this method removes the gaps between the lower oracle complexity bounds of the function class. Furthermore, the Nesterov method corrects the gradient in every iteration and makes the update of the gradient more flexible, in some way, which can accelerate the convergence.

2.5 The TV-AB algorithm

We now briefly introduce the existing method to solve problem (1). Fakhteh et al. [27] provide a novel method and the distributed form of the proposed algorithm is given as follows:

$$x_{k+1}^i = \sum_{j=1}^{m} A_{ik}^j x_j^k - \alpha y_k^i, \quad (8a)$$

$$y_{k+1}^i = \sum_{j=1}^{m} B_{ik}^j y_j^k + \nabla f_i (x_{k+1}^i) - \nabla f_i (x_k^i), \quad (8b)$$

where $A_{ik}^j$ and $B_{ik}^j$ are row- and column-stochastic matrices, respectively, and $\alpha$ is the constant step-size. At each iteration, agent $i \in \mathcal{V}$ maintains two variables: $x_{k+1}^i, y_k^i \in \mathbb{R}^n$, where $x_{k+1}^i$ is the local estimate of the optimal solution and $y_k^i$ is an estimate of the global gradient. We can clearly see that the update (8a) is basically gradient descent with the direction controlled by $y_k^i$ instead of the local gradient $\nabla f_i (x_k^i)$. Evidently, the update (8b) tracks the average of local gradients, $\frac{1}{m} \sum_{i=1}^{m} \nabla f_i (x_k^i)$, referring to [26–31, 39]. When dealing with the distributed optimisation problem over time-varying directed networks, TV-AB algorithm converges linearly. However, an explicit range of its step-size fails to be derived and the accelerated techniques are not considered.

2.6 Distributed algorithm development

Motivated by the TV-AB algorithm [27], we provide a novel algorithm that incorporates distributed heavy-ball method and distributed Nesterov method into the iteration. At iteration $k$, each agent $i \in \mathcal{V}$ controls three variables: $x_{k+1}^i, s_{k+1}^i, y_k^i \in \mathbb{R}^n$, initialised with arbitrary $x_0^i = x_0^i, s_0^i = s_0^i$, and $y_0^i = \nabla f_i (x_0^i)$. The distributed form of the proposed algorithm is given as follows:

$$x_{k+1}^i = \sum_{j=1}^{m} A_{ik}^j x_j^k - \alpha y_k^i + \beta_i (x_k^i - \sum_{j=1}^{m} A_{ik}^j x_{k-1}^j), \quad (9a)$$

$$s_{k+1}^i = x_{k+1}^i + \beta_i (s_{k+1}^i - \sum_{j=1}^{m} A_{ik}^j s_{k-1}^j), \quad (9b)$$

$$y_{k+1}^i = \sum_{j=1}^{m} B_{ik}^j y_j^k + \nabla f_i (s_{k+1}^i) - \nabla f_i (y_k^i), \quad (9c)$$

where the weights $A_{ik}^j$ and $B_{ik}^j$ satisfy Assumption 3. Both the uncoordinated step-size, $\alpha$, and the momentum parameter, $\beta_i$, are locally decided by agent $i$. As is shown in the proposed algorithm, we incorporate the heavy-ball method [28] into (9a), which can be nearly seen as a centralised gradient descent. In (9b), we bring in another type of acceleration that is the distributed Nesterov method [31]. By using the outcome of update (9b), (9c) is performed as the gradient-tracking update.

We now write the matrix-vector form of (9a)–(9c). The local variables $x_k^i, s_k^i, y_k^i$ are collected by $x_k, s_k, y_k \in \mathbb{R}^{mn}$, respectively, and $\nabla F (s_k) = [\nabla f_1^T (s_1), \ldots, \nabla f_m^T (s_m)]^T$. Define $A_k = A_k \otimes I_m$, $B_k = B_k \otimes I_m$, $D_k = \text{diag}(\alpha) \otimes I_m$. Furthermore, $\gamma_k = \text{diag}(\beta) \otimes I_m$, where $\alpha$ is the vector of local step-sizes and $\beta$ is the vector of momentum parameters. Then, initialised with arbitrary $x_0, s_0, y_0 = \nabla F (s_0)$, the matrix-vector form of (9) is briefly summarised in the following:

$$x_{k+1} = A_k y_k + D_k y_k - D_k \gamma_k, \quad (10a)$$

$$s_{k+1} = x_{k+1} + D_k (x_{k+1} - A_k y_k), \quad (10b)$$

$$y_{k+1} = B_k y_k + \nabla F (s_{k+1}) - \nabla F (s_k), \quad (10c)$$

3 CONVERGENCE ANALYSIS

In what follows, we start with a state transformation: $g_k = (V_{k-1}^{-1} \otimes I_m) g_k$, where $V_k = \text{diag}(v_k)$ and $v_k$ follows (11a). Thus, algorithm (10) can be equivalently rewritten as follows:

$$v_{k+1} = B_k v_k \quad (11a)$$

$$x_{k+1} = A_k x_k + D_k (A_k + I_m) x_k - D_k \gamma_k, \quad (11b)$$

$$s_{k+1} = x_{k+1} + D_k (x_{k+1} - A_k x_k), \quad (11c)$$

$$g_{k+1} = R_k g_k + (V_{k-1}^{-1} \otimes I_m) (\nabla F (s_{k+1}) - \nabla F (s_k)), \quad (11d)$$

where $R_k = R_k \otimes I_m$, $R_k = V_{k-1}^{-1} B_k V_k$ and $v_0 = 1_m$. One can testify that $\{R_k\}$ is a sequence of row-stochastic matrices with the absolute probability sequence, $\{v_k\}$. The definition of absolute probability sequence is presented in the next.

Definition 1 ([40]). For row-stochastic matrices, $\{R_k\}$, an absolute sequence is a sequence $\{\pi_{k}\}$ of stochastic vectors such that

$$\pi_k^T = \pi_{k+1}^T R_k, \forall k \geq 0.$$ 

Definition 2 ([41]). An ergodic sequence of row-stochastic matrices, $\{R_k\}$, is such that for integers $n \geq 0$ and all $i \in \mathcal{V}$, $s = 1, \ldots, m$

$$\lim_{i \to \infty} \left| U_{(s)} \right| = d_n^s,$$

where $U_{(s)} = \prod_{i=0}^{n-1} R_i$ is the backward product of $\{R_i\}$ and $d_n^s$ is a constant independent of $i$. 


We next analyse the convergence of algorithm (11). Our analysis depends on the quantities described as follows: (i) \( \hat{x}_k^w = (\hat{x}_k^w \otimes I_c)\) \( x_k^w \), which is the average of \( x_k^w \) weighted by the absolute probability sequence \( \{\mu_k^w\} \) of \( A_k \); (ii) The networked weighted consensus error is set as: \( \hat{x}_k^w = x_k - 1_m \otimes \hat{x}_k^w \); (iii) The optimal gap corresponding to the weighted estimation can be built as: \( r_k = 1_m \otimes \hat{x}_k^w - 1_m \otimes \hat{x}_k \); (iv) \( d_k = x_k - A_{k-1}x_{k-1} \), which can be regarded as state difference in consecutive update time; (v) An error term associated with gradient estimation is represented by \( \hat{x}_k^w = \hat{x}_k - (1_m \otimes I_c)\). Leveraging the above quantities, we define the vector, \( t_k := \|\hat{x}_k^w\|, \|r_k\|, \|d_k\|, \|\hat{x}_k^w\|^2 \), to prove \( t_k \to 0 \) when \( k \to \infty \). Clearly, we obtain that \( x_k \) converges to \( x^* \) if \( t_k \to 0 \). For the following analysis, we aim at establishing a system inequality as follows:

\[
\begin{bmatrix}
    t_{k+1} \\
    \vdots \\
    t_{k-\ell+2} \\
    t_k \\
    \vdots \\
    t_{k-\ell+1} 
\end{bmatrix} \leq \begin{bmatrix}
    M_\alpha \beta \\
    \vdots \\
    M_\alpha \beta \\
    \vdots \\
    M_\alpha \beta 
\end{bmatrix} 
\]

where \( \alpha = [\alpha_{\max}, \beta = [\beta_{\max}] \) and the coefficients of the linear system are the elements of \( M_\alpha \beta \). Clearly, if \( \rho(M_\alpha \beta) < 1 \), then \( x_k \) converges to \( x^* \) at least at the rate of \( \Theta(\rho(M_\alpha \beta)) \). Note that the constant \( \max(C_A \tilde{C}_B) \) guarantees concurrent multi-step contractions for \( \hat{x}_k^w \) and \( \hat{x}_k \), where \( \tilde{C}_A \) and \( \tilde{C}_B \) refer to Lemma 2 and Corollary 2, respectively. In the following, we present some necessary lemmas.

### 3.1 Auxiliary relations

**Lemma 1 ([27]).** Suppose that Assumptions 2 and 3 hold. The row-stochastic matrix sequence \( \{D_s\} = \prod_{i=s+1}^{s+C-1} A_i \) compliant with the aggregate directed network \( G^w = (\Psi, \cup_{s=\ell}^{s+C-1} E_s) \), for all \( s \geq 0 \), is ergodic, that is

\[
\lim_{k \to \infty} D_s \ldots D_{s+1}D_s = 1_m \mu^T, 
\]

where \( \{\mu_s\} \) is the unique absolute probability sequence for \( \{D_s\} \) and is uniformly bounded away from zero, that is, there exists \( \delta \in (0, 1) \) such that \( \mu_s \geq \delta \) \( \forall s \geq 0 \). The convergence rate is indeed geometric, that is, \( t_k \geq s \geq 0 \):

\[
\|D_s \ldots D_{s+1}D_s - 1_m \mu^T\| \leq Kq^{s-k},
\]

where the constants \( K > 0 \) and \( q \in (0, 1) \) rely on \( w_1 \) and \( m \) in Assumption 3.

The next corollary derives the absolute probability sequence for the sequence \( \{A_k\} \) in terms of \( \{\mu_k\} \) in Lemma 1.

**Corollary 1.** Suppose that Assumptions 2 and 3 hold. Then, the sequence \( \{\phi_k\} \) is an absolute probability sequence for the matrix sequence \( \{A_k\} \), where \( \phi_k^T = \mu_k^T, k = k \) \( C \) and for some \( s \geq 0, \phi_k^T \) is given by

\[
\phi_k^T = \mu_{(s+1)C-1}^T A_{(s+1)C-1} \ldots A_k, k \in (sC, (s+1)C).
\]

**Proof.** From Lemma 1, we know that \( \{\mu_k\} \) is an absolute probability sequence and the products \( A_{(s+1)C-1} \ldots A_k \) are row-stochastic. Both \( \{\mu_k\} \) and \( \{\phi_k\} \) are stochastic vectors, then

\[
\phi_k^T = \mu_{(s+1)C-1}^T A_{(s+1)C-1} \ldots A_k
\]

\[
= \mu_{(s+1)C-1}^T A_{(s+1)C-1} \ldots A_{k+1} A_{k}
\]

\[
= \phi_{k+1}^T A_k.
\]

We next construct \( \tilde{C}_A \)-step contraction for \( \{A_k\} \), which is fundamental for the convergence analysis of the proposed algorithm.

**Lemma 2 ([22]).** Suppose that Assumptions 2 and 3 hold. Recall \( A_k = A_k \otimes I_c \) such that

\[
y_A = Q_A(1 - \alpha^\infty C_A - \infty < 1,
\]

where \( Q_A = \begin{bmatrix} 2m^{1+\varepsilon C} & 2m^{1+\varepsilon C} \end{bmatrix} \) and \( C_A \geq C \) is an integer. For any vector \( b \in \mathbb{R}^m \), if \( a = A_k, \tilde{C}_A \) and \( A_k, \tilde{C} = A_k A_{k-1} \ldots A_{k-\tilde{C}+1} \), it holds

\[
\|((I_m - 1_m \phi^T_{k+1} \otimes I_c) b)\| \leq y_A \|((I_m - 1_m \phi^T_{k-\tilde{C}+1}) \otimes I_c) b)\|,
\]

where \( \{\phi_k\} \) defined in Corollary 1 is an absolute probability sequence of \( \{A_k\} \).

The next corollary constructs the multi-step contraction for the sequence \( R_k = V_{k+1}^T B_k V_k \), where \( V_k = \text{diag}(v_k) \).

**Corollary 2 ([27]).** Suppose that Assumptions 2 and 3 hold. Recall \( R_k = R_k \otimes I_c \) such that

\[
y_B = Q_B(1 - \alpha^\infty C_B - \infty < 1,
\]

where \( Q_B = \begin{bmatrix} 2m^{1+\varepsilon C} & 2m^{1+\varepsilon C} \end{bmatrix} \) and \( C_B \geq C \) is an integer. Then, for \( k \geq \tilde{C}_B - 1 \) and any vector \( b \in \mathbb{R}^m \), if \( a = R_k, \tilde{C}_B \), we have

\[
\|((I_m - R_k^T \phi) b)\| \leq y_B \|((I_m - R_k^T \phi_{k-\tilde{C}_B+1}) \otimes I_c) b)\|.
\]

The following lemma is not uncommon in the convex optimisation theory [26, 28], which shows a fixed ratio of the contraction to optimal solution.

**Lemma 3 ([28]).** Let \( f \) be \( L_f \)-smooth and \( \mu \)-strongly convex. For any \( \tilde{x} \in \mathbb{R}^n \) and \( 0 < \delta < \frac{1}{L_f} \), it holds

\[
\|\tilde{x} - \delta\nabla f(\tilde{x}) - \tilde{x}^*\| \leq \lambda \|\tilde{x} - \tilde{x}^*\|,
\]

where \( \lambda = 1 - \mu\delta \).

To proceed, we start with seeking an upper bound on \( y_k \).
Lemma 4. Suppose that Assumption 1 holds. For \( \forall k \geq 0 \), we have an inequality holds as follows:

\[
\|y_t\| \leq mL_f \|\tilde{x}_{k+1}^p\| + mL_f \|r_k\| + \tilde{\beta} mL_f \|d_k\| + \|\tilde{z}_{k+1}^p\|.
\]

Proof. See Appendix A.

Lemma 5. Suppose that Assumptions 1–3 hold. For \( \forall k \geq 0 \), we have an inequality holds as follows:

\[
\|	ilde{z}_{k+1}^p\|
\leq \sum_{i=0}^{\tilde{c}-1} \left[ \|\mathcal{A}_k + I_m\| \|\tilde{z}_{k-i}^p\| + \alpha mL_f \|r_{k-i}\| + \tilde{\beta} (\|\mathcal{A}_k + I_m\| + \alpha mL_f) \|d_{k-i}\| + \gamma \tilde{\beta} \|z_{k-i}\| \right]
\]

Proof. See Appendix B.

Lemma 6. Suppose that Assumptions 1–3 hold. For \( \forall k \geq 0 \), we have an inequality holds as follows:

\[
\|r_{k+1}\| \leq \tilde{\beta} \left( \|\mathcal{A}_k + I_m\| \|\tilde{z}_{k}^p\| + \alpha mL_f \|r_k\| + \tilde{\beta} mL_f \|d_k\| \right)
\]

Proof. See Appendix C.

Lemma 7. Suppose that Assumptions 1–3 hold. For \( \forall k \geq 0 \), we have an inequality holds as follows:

\[
\|d_{k+1}\| \leq \tilde{\beta} \left( \|\mathcal{A}_k + I_m\| \|\tilde{z}_{k}^p\| + \alpha mL_f \|r_k\| + \tilde{\beta} mL_f \|d_k\| \right)
\]

Proof. See Appendix D.

Lemma 8. Suppose that Assumptions 1–3 hold. For \( \forall k \geq 0 \), we have an inequality holds as follows:

\[
\|	ilde{z}_{k+1}^p\|
\leq \sum_{i=0}^{\tilde{c}-1} \left[ \|\mathcal{A}_k + I_m\| \|\tilde{z}_{k-i}^p\| + \alpha mL_f \|r_{k-i}\| + \tilde{\beta} mL_f \|d_{k-i}\| \right]
\]

where \( t = \frac{1}{L_f \max \{Q_a\}} > 0 \).

Proof. See Appendix E.

Lemma 9 ([42]). For a nonnegative matrix, \( X \in \mathbb{R}^{\infty \times \infty} \), and an arbitrary positive vector, \( x \in \mathbb{R}^n \). We have \( \rho(X) < w \) when \( X x < w x \) with \( w > 0 \).

We next define some key matrices used in Theorem 1 as follows:

\[
W_i = \begin{bmatrix}
\tilde{\omega}_1 & \tilde{\omega}_5 & 1 - \tilde{\omega}_4 & 2\tilde{\omega}_5 + 3\tilde{\omega}_3 & \tilde{\omega}_3 \\
\tilde{\omega}_3 & \tilde{\omega}_4 & \tilde{\omega}_5 & \tilde{\omega}_3 & \tilde{\omega}_3 \\
\tilde{\omega}_5 & \tilde{\omega}_6 & \tilde{\omega}_7 & \tilde{\omega}_8 & \tilde{\omega}_9 \\
\end{bmatrix}
\]

\[
W_j = \begin{bmatrix}
\gamma + \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

where the definitions of parameters \( \omega_i \) and \( W_i \) are placed in Appendix F for clarity of exposition.

Lemma 10 ([42]). Suppose that there are two nonnegative matrices with some dimensions, \( T \) and \( \tilde{T} \). If any element of \( \tilde{T} \) is no less than the counterpart element in \( T \) such that \( T^{ij} \leq \tilde{T}^{ij} \), then we have \( \rho(T) \leq \rho(\tilde{T}) \).

Therefore, we redefine aforementioned matrices as follows:

\[
W_i = \begin{bmatrix}
\tilde{\omega}_1 & \tilde{\omega}_1 & 1 - \tilde{\omega}_4 & 2\tilde{\omega}_5 + 3\tilde{\omega}_3 & \tilde{\omega}_3 \\
\tilde{\omega}_3 & \tilde{\omega}_4 & \tilde{\omega}_5 & \tilde{\omega}_3 & \tilde{\omega}_3 \\
\tilde{\omega}_5 & \tilde{\omega}_6 & \tilde{\omega}_7 & \tilde{\omega}_8 & \tilde{\omega}_9 \\
\end{bmatrix}
\]

\[
W_j = \begin{bmatrix}
\gamma + \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 & \tilde{\omega}_1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
where the parameters $\tilde{\Omega}_i$ and $\tilde{\Omega}_i'$ are defined in Appendix F. Furthermore, for simplifying the presentation of the calculation, we define some parameters used in Theorem 1 as follows:

$$
\Omega_1 = \omega_1 \theta_{11} + \omega_1 \theta_{12} + \omega_1 \theta_{14} + \sum_{i=2}^{C-1} \omega_i \theta_{i1} + \omega_i \theta_{i2} + \omega_i \theta_{i4},
$$

$$
\Omega_2 = \omega_2 \theta_{11} + \omega_2 \theta_{13} + \omega_2 \theta_{14} + \sum_{i=2}^{C-1} \omega_i \theta_{i1} + \omega_i \theta_{i3} + \omega_i \theta_{i4},
$$

$$
\Omega_3 = (\omega_6 + \alpha \omega_1) \theta_{13} + \sum_{i=2}^{C-1} (\omega_i + \alpha \omega_1) \theta_{i3} + (\omega_i + \alpha \omega_1) \theta_{i3},
$$

$$
\psi_1 = (\omega_1 + \alpha \omega_8) \theta_{12} + \sum_{i=2}^{C-1} (\omega_i + \alpha \omega_8) \theta_{i2} + (\omega_i + \alpha \omega_8) \theta_{i2},
$$

$$
\psi_2 = \sum_{i=2}^{C-1} (39 \omega_8 \theta_{i1} + (\omega_8 + \alpha \omega_8) \theta_{i1} + \alpha \omega_8 \theta_{i3} + \alpha \omega_8 \theta_{i4})
$$

$$
\psi_3 = \sum_{i=2}^{C-1} ((\omega_i + \alpha \omega_8) \theta_{i1} + \alpha \omega_i \theta_{i3} + \alpha \omega_i \theta_{i4})
$$

$$
\sum_{i=2}^{C-1} ((\omega_i + \alpha \omega_i) \theta_{i1} + \alpha \omega_i \theta_{i3} + \alpha \omega_i \theta_{i4})
$$

$$
\sum_{i=2}^{C-1} \theta_{i1} + \theta_{i3} + \theta_{i4}.
$$

\subsection{3.2 Main results}

\textbf{Theorem 1.} Suppose that Assumptions 1-3 hold and let $0 < \tilde{\alpha} < \frac{1}{\max_{i,j} \Omega_i}$. Then for all $k \geq C-1$, the following inequality holds

$$
t_k + 1 \leq \tilde{\Omega}_k t_k + \sum_{j=2}^{C-1} \tilde{\Omega}_j t_{j-k-1} + \tilde{\Omega}_k t_{k-C-1} + 1,
$$

which is equivalent to

$$
\begin{bmatrix} t_k+1 \\ t_k \\ \vdots \\ t_{k-C+2} \end{bmatrix} = \begin{bmatrix} W_1 \cdots W_{C-1} \\ I \cdots I \end{bmatrix} \begin{bmatrix} t_k \\ t_{k-1} \\ \vdots \\ t_{k-C+1} \end{bmatrix} = T_{\tilde{\alpha}, \tilde{\theta}} \begin{bmatrix} t_k \\ t_{k-1} \\ \vdots \\ t_{k-C+1} \end{bmatrix}.
$$

Defining positive vectors $\theta_i = [\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4}]^T, i = 1, 2, \ldots, C$, if

$$
\tilde{\Omega}_1 \theta_{11} + \sum_{i=2}^{C-1} \tilde{\Omega}_i \theta_{i1} + \tilde{\Omega}_C \theta_{C} < \theta_1
$$

and $\theta_i < \theta_{i+1}, i = 1, 2, \ldots, C-1,$ we obtain that $\rho \left( \tilde{\Omega}, \tilde{\theta} \right) < 1$. Thus, $\| x_k - x^* \|$ converges linearly to zero at least at the rate of $\rho \left( \tilde{\Omega}, \tilde{\theta} \right)^k$ when the largest step-size satisfies

$$
0 < \tilde{\alpha} < \min
$$

$$
\begin{bmatrix} \frac{\Omega_2 \theta_{11} - \theta_{13}}{\Omega_1} \\ \frac{\Omega_2 \theta_{12} + \omega_1 \theta_{12} + \theta_{14}}{\Omega_2} \end{bmatrix} < \tilde{\beta}
$$

and the largest momentum parameter satisfies

$$
0 \leq \tilde{\beta} < \min
$$

$$
\begin{bmatrix} \frac{1}{\Omega_1} [\theta_{11} - \tilde{\alpha} \Omega_1 - \gamma_4 \theta_{C1}^T] \\ \frac{\Omega_1}{\Omega_1} [\theta_{13} - \tilde{\alpha} (\omega_2 \theta_{12} + \omega_3 \theta_{13} + \theta_{14})] \end{bmatrix} < \tilde{\beta}
$$

$$
\begin{bmatrix} \frac{1}{\Omega_2} [\theta_{14} - \tilde{\alpha} \theta_{11} - \tilde{\alpha} \theta_{13} - \gamma_4 \theta_{C4}^T] \\ \frac{\Omega_2}{\Omega_2} [\theta_{13} - \tilde{\alpha} (\omega_2 \theta_{12} + \omega_3 \theta_{13} + \theta_{14})] \end{bmatrix} < \tilde{\beta}
$$

\textbf{Proof.} Leveraging the results from Lemmas 5–8, one can directly verify (16). Note that $\rho \left( \tilde{\Omega}, \tilde{\theta} \right) \leq 0$, if $\rho \left( \tilde{\Omega}, \tilde{\theta} \right) < 1$, we obtain that $\rho \left( \tilde{\Omega}, \tilde{\theta} \right) < 1$. The next step aims to find the ranges of $\tilde{\alpha}$ and $\tilde{\beta}$ such that $\rho \left( \tilde{\Omega}, \tilde{\theta} \right) < 1$. Based on Lemma 9, we solve for positive vectors $\theta_i, i = 1, 2, \ldots, C$, such that the ranges of $\tilde{\alpha}$ and $\tilde{\beta}$ can be obtained by resolving the inequality as follows:

$$
\tilde{\Omega}_i \theta_i < \theta_i,
$$

where $\theta = [\theta_1^T, \theta_2^T, \ldots, \theta_C^T]^T$. It can be equivalently rewritten as follows:

$$
\begin{bmatrix} \frac{\Omega_1}{\Omega_2} \theta_1 + \sum_{i=2}^{C-1} \Omega_i \theta_{i1} + \Omega_C \theta_{C} < \theta_1 \\ \theta_2 < \theta_2 \end{bmatrix}
$$

To proceed, we can select some positive vectors, $\theta_i, i = 1, 2, \ldots, C-1$, such that $\theta_i < \theta_{i+1}$. Then, the ranges of $\tilde{\alpha}$ and $\tilde{\beta}$
can be obtained by resolving the inequality as follows:

\[ \tilde{\psi}_1 \theta_1 + \sum_{j=2}^{\tilde{c}-1} \tilde{\psi}_j \theta_j + \tilde{\psi}_C \theta_C < \theta_1, \]  

(20)

which is equivalent to

\[ \bar{\beta} \Omega_3 < \theta_{11} - \bar{\alpha} \Omega_1 - \gamma_A \tilde{\theta}_1, \]  

(21)

\[ \bar{\beta} (2 \omega_3 + \bar{\alpha} \omega_3) \theta_{13} < \bar{\alpha} (\omega_3 \theta_{12} - \omega_3 \theta_{11} - \omega_3 \theta_{14}), \]  

(22)

\[ \bar{\beta} (\omega_6 + \bar{\alpha} \omega_3) \theta_{13} < \theta_{13} - \bar{\alpha} (\omega_3 \theta_{12} + \omega_3 \theta_{14} + \theta_{11}), \]  

(23)

\[ \psi_1 \bar{\beta}^2 + \psi_2 \bar{\beta} < \psi_3, \]  

(24)

where \( \Omega_3, i = 1, 2, 3, \) and \( \psi_i, i = 1, 2, 3, \) are defined before Theorem 1. In order to enable the right-hand-side of inequalities (21)–(24) to be positive, the following inequalities should be satisfied.

\[
\begin{align*}
0 < \bar{\alpha} &< \frac{\theta_{11} - \gamma_A \tilde{\theta}_1}{\Omega_1}, \\
0 &< \gamma_A \tilde{\theta}_1 < \theta_{11}, \\
0 &< \bar{\alpha} < \frac{\omega_3 \theta_{11} + \omega_3 \theta_{14}}{\omega_3}, \\
0 &< \bar{\alpha} < \frac{\theta_{13}}{\omega_3 \theta_{11} + \omega_3 \theta_{12} + \theta_{14}}. 
\end{align*}
\]

(25)

(26)

(27)

One can pick an arbitrary positive constant \( \theta_{13} \) at first, and next choosing \( \theta_{11} \) satisfies (25). Finally, choosing \( \theta_{14} \) satisfies (28) and picking \( \theta_{12} \) satisfies (26). Then, summarising (25), (26), (27) and (28), together with the precondition: \( 0 < \bar{\alpha} < \frac{1}{n \lambda_f} \), we get the exact range of \( \bar{\alpha} \) in (17) while the exact range of \( \bar{\beta} \) in (18) can be obtained directly from (21), (22), (23), and (24). Recalling the definition of \( \ell_k \), it is not difficult to verify that \( \| x_k - x^* \| \) converges linearly to zero at least at the rate of \( \mathcal{O}(\ell_k^{k^2}) \).

**Remark 2.** It is worth mentioning that the theoretical bounds on the largest step-size and the largest momentum parameter derived in Theorem 1 are conservative. In fact, one can manually choose more appropriate step-sizes and momentum parameters in practice. We also find that the bounds on the largest step-size and the largest momentum parameter contain some global parameters, such as \( \sigma_A \) and \( \sigma_B \). Therefore, the use of sufficiently small but positive step-sizes and momentum parameters to theoretically guarantee all the communication and computation process conducted in a fully local way, which is not uncommon in the recent literature [23, 28]. Besides, we emphasize that the theoretical analysis on the acceleration developed by the distributed heavy-ball method and distributed Nesterov method remains an open problem. This paper demonstrates the acceleration through a series of simulations.

In following sections, we apply the proposed algorithm into fixed directed networks, which can be considered as a special case of the time-varying directed networks. Specifically, in fixed directed networks, we assume that \( A_k = A, B_k = B, A_k = A, B_k = B, k = 0, 1, \ldots \). The next assumption serves as a special case of Assumption 2.

**Assumption 4.** The directed network, \( G \), is fixed strongly connected.

**Remark 3.** Assumption 4 indicates that the positive integer \( C = 1 \) in time-varying networks. Define \( A = A \otimes I_m, B = B \otimes I_m, A_\infty = 1 \pi_m^T, \) and \( B_\infty = \pi_B I_m^\perp \) where \( \pi_i \in \mathbb{R}^m \) is the left eigenvector of a row-stochastic matrix, \( A \), and \( \pi_i \in \mathbb{R}^m \) is the right eigenvector of a column-stochastic matrix, \( B \). The next lemma is built on the above assumptions.

**Lemma 11 ([32]).** Suppose that the communication network, \( G \), is fixed strongly connected. Considering the augmented weight matrices, \( A \) and \( B \), there exists vector norms, \( \| \cdot \|_{\pi_r} \) and \( \| \cdot \|_{\pi_c} \), such that

\[
\| A \pi_r - A_\infty \pi_r \|_{\pi_c} \leq \sigma_A \| \pi_c - A_\infty \|_{\pi_c},
\]

\[
\| B \pi_r - B_\infty \pi_r \|_{\pi_c} \leq \sigma_B \| \pi_c - B_\infty \|_{\pi_c},
\]

where \( 0 < \sigma_A \triangleq \| A - A_\infty \|_{\pi_c} < 1 \) and \( 0 < \sigma_B \triangleq \| B - B_\infty \|_{\pi_c} < 1 \) are some constants.

Before presenting Theorem 2, we first define an important matrix as follows:

\[
W = \begin{bmatrix}
\sigma_A + \bar{\alpha} \omega_17 & \bar{\alpha} \omega_17 & \tilde{\beta} \omega_18 + \bar{\alpha} \tilde{\beta} \omega_17 & \bar{\omega}_9 \\
\bar{\alpha} \omega_20 & 1 - \bar{\alpha} \omega_21 & \bar{\alpha} \tilde{\beta} \omega_20 + \bar{\alpha} \tilde{\beta} \omega_22 & \bar{\alpha} \omega_23 \\
\bar{\alpha} \omega_24 & \bar{\alpha} \omega_24 & \tilde{\beta} \omega_25 + \bar{\alpha} \tilde{\beta} \omega_24 & \bar{\alpha} \\
\mathcal{W}_8 & \bar{\omega}_8 + \bar{\alpha} \tilde{\beta} \omega_8 & \mathcal{W}_9 & \mathcal{W}_{10}
\end{bmatrix},
\]

where \( \sigma_A, \sigma_B \) are defined in Lemma 11 and the definitions of parameters \( \mathcal{W}_i, i = 8, 9, 10, \) and \( \omega_i, i = 17, \ldots, 29 \) are
placed in Appendix F. Due to the introduction of the vector norms, \( \| \cdot \|_{\pi_r} \) and \( \| \cdot \|_{\pi_\tau} \), we need to define a new vector \( \tilde{t}_k = [\|x_k\|_{\pi_r}, \|r_k\|, \|d_k\|, \|x_k\|_{\pi_\tau}]^T \). Moreover, we define some parameters used in Theorem 2 as follows:

\[
\psi_4 = (\omega_{28} + \bar{\kappa}_2 \omega_{28}) \theta_3,
\psi_5 = \bar{\kappa}_2 \omega_{28} \theta_1 + \bar{\kappa}_2 \omega_{28} \theta_3 + (\omega_{28} + \omega_{29} + \bar{\kappa}_2 \omega_{28}) \theta_3 + \bar{\kappa}_2 \omega_{29} \theta_4,
\psi_6 = (1 - \sigma_B - \bar{\kappa}_2 \omega_{28}) \theta_4 - (\omega_{26} + \bar{\kappa}_2 \omega_{27}) \theta_1 - \bar{\kappa}_2 \omega_{27} \theta_2.
\]

**Theorem 2.** Suppose that Assumptions 1, 4 hold and let \( 0 < \bar{\kappa} < \frac{1}{mL_f} \). Then for \( \forall k \geq 0 \), the following inequality holds:

\[
\tilde{t}_{k+1} \leq W \tilde{t}_k.
\]

When the largest step-sizes, \( \bar{\kappa} \), satisfies

\[
0 < \bar{\kappa} < \min \left\{ \frac{\theta_1 + \bar{\kappa}_2 \omega_{28} \theta_1 + \omega_{28} \theta_2 + \theta_4}{(1 - \sigma_A) \theta_1 - (\omega_{28} + \bar{\kappa}_2 \omega_{28}) \theta_3}, \frac{(1 - \sigma_B) \theta_1 - (\omega_{28} + \bar{\kappa}_2 \omega_{28}) \theta_3}{\omega_{27} \theta_1 + \omega_{25} \theta_2 + \omega_{29} \theta_4}, \frac{1}{mL_f} \right\},
\]

and the largest momentum parameter, \( \bar{\beta} \), satisfies

\[
0 \leq \bar{\beta} < \min \left\{ \frac{\theta_1 - \bar{\kappa}_2 \omega_{28} \theta_1 - \bar{\kappa}_2 \omega_{28} \theta_3 - \bar{\kappa}_2 \omega_{29} \theta_4}{(\omega_{19} + \bar{\kappa}_2 \omega_{17}) \theta_3}, \frac{(1 - \omega_{21}) \theta_2 - \bar{\kappa}_2 \omega_{24} \theta_2 - \bar{\kappa}_2 \omega_{23} \theta_4}{(\omega_{23} + \bar{\kappa}_2 \omega_{23} \theta_4)}, \frac{\theta_1 - \bar{\kappa}_2 \omega_{24} \theta_2 - \bar{\kappa}_2 \omega_{24} \theta_3 - \bar{\kappa}_2 \omega_{23} \theta_4}{\omega_{25} + \bar{\kappa}_2 \omega_{25} \theta_4}, \frac{-\psi_5 + \sqrt{\psi_5^2 + 4\psi_4 \psi_6}}{2\psi_4} \right\},
\]

where \( \theta_1, \theta_2, \theta_3, \theta_4 \) are arbitrary constants such that

\[
0 < \theta_1 < \frac{(1 - \sigma_B) \theta_4}{\omega_{26}},
\theta_2 > \frac{\omega_{21} \theta_1 + \omega_{23} \theta_4}{\omega_{21}} > 0,
\theta_3 > 0,
\theta_4 > 0
\]

then \( p(W) < 1 \) and thus \( \|x_k - x^*\| \) converges linearly to zero at least at the rate of \( O(p(W)^{k}) \).

**Proof.** The proofs are in line with that of in Theorem 1 and thus the detailed proofs are omitted here for length considerations.

---

**4 | SIMULATION RESULTS**

In order to illustrate the performance and applications of the proposed algorithm, some numerical simulations are conducted in this section. We utilise a simple uniform weighting tactics to construct row- and column-stochastic weights \( A_k^f, B_k^f \) :

\[
A_k^f = \begin{cases} 
\frac{1}{d_{k,in}^f}, & (j, i) \in E_k \\
0, & (j, i) \notin E_k 
\end{cases},
B_k^f = \begin{cases} 
\frac{1}{d_{k,\text{out}}^f}, & (j, i) \in E_k \\
0, & (j, i) \notin E_k 
\end{cases},
\]

where \( d_{k,in}^f \) is the in-degree of agent \( i \) at iteration \( k \) and \( d_{k,\text{out}}^f \) is the out-degree of agent \( j \) at iteration \( k \). The performance of all the tested algorithms are described by plotting the residual, \( \frac{1}{m} \sum_{i=1}^{m} \|x_k^i - \bar{x}^i\| \). In the following simulations, we assume that the samples are distributed equally over the agents, that is, \( q_i = N/m, i \in V \), where \( N \) is the total number of samples and \( m \) represents the number of agents in the network. Recall that \( \bar{\kappa} \) and \( \bar{\beta} \) denote the largest constant step-size, \( \kappa_\alpha \), and the largest momentum parameter, \( \beta_\alpha, i = 1, \ldots, m \), respectively. We emphasise that in the following simulations, EXTRA works in a fixed undirected network which is generated by the Laplacian method [28].

**4.1 | Distributed least-squares**

This simulation considers a least-squares problem provided in [3] that solves for an unknown signal \( \bar{x} \in \mathbb{R}^n \), where a network of \( m = 5 \) agents cooperatively resolve the following optimisation problem

\[
\min_{\bar{x} \in \mathbb{R}^n} \frac{1}{2} \| C \bar{x} - d \|^2,
\]

where \( C_i \in \mathbb{R}^{q_i \times n} \) is the sensing matrix; \( d_i = C_i \bar{x} + e_i \) is the measurement given by agent \( i = 1, \ldots, 5 \) and \( e_i \in \mathbb{R}^{q_i} \) is the independent and identically distributed noise. We set \( q_i = 1000 \) and the dimensions of all signals are assumed to be \( n = 10 \), while each dimension represents one feature of the signals. As is shown in Figure 1, \( m = 5 \) agents communicate with each other over a time-varying network with \( C = 4 \). Figure 2 demonstrates that the proposed algorithm shows better performance than the other tested algorithms and the convergence rates are linear.

**4.2 | Distributed quadratic programming**

In this case study, we aim at exploring the impact on the acceleration brought by the changes of condition number \( Q = L_f/\mu \). Assume that a network of \( m = 50 \) agents cooperatively resolve
a quadratic programming problem as follows:

$$
\min_{\tilde{x} \in \mathbb{R}^n} \tilde{f}(\tilde{x}) = \sum_{i=1}^{m} \tilde{x}^T G_i \tilde{x} + c_i^T \tilde{x},
$$

where matrices $G_i \in \mathbb{R}^{n \times n}$ are diagonal and positive-definite; vectors $c_i \in \mathbb{R}^n$ are randomly generated; the dimension is set as $n = 100$. Figure 3 depicts a time-varying directed network at a particular moment. The condition number, $Q$, of $\tilde{f}$ is calculated by the ratio of the largest to the smallest eigenvalue of $G = \sum_{i=1}^{m} G_i$. It is worth mentioning that the proposed algorithm reduces to TV-$\mathcal{AB}$ with uncoordinated stepsizes when the momentum parameter $\beta = 0$. From Figures 4–6, one can clearly see that the proposed algorithm with accel-
ated momentum outperforms the proposed algorithm without acceleration and the other tested algorithms when the condition number, $\mathcal{Q}$, changes. Note that we have to choose the appropriate step-sizes to ensure the convergence of the tested algorithm under different condition number $\mathcal{Q}$.

4.3 Distributed logistic regression

In this experiment, large-scale binary classification problems based on real-world data set are solved by the proposed algorithm and other tested algorithms over a fixed directed network. The global objective function can be formulated as $\tilde{f}(\tilde{x}) = \sum_{i=1}^{m} f_i(\tilde{x})$ and the local loss objective function, $f_i$, can be formulated as:

$$f_i(\tilde{x}) = \frac{\beta}{2m} \|\tilde{x}\|^2 + \sum_{j=1}^{q_i} \log \left(1 + \exp \left(-b_{ij} c_{ij}^T \tilde{x}\right)\right),$$

where $c_{ij} \in \mathbb{R}^n$ is the $j$-th training sample and $b_{ij} \in \{+1, -1\}$ is the corresponding label, both of which are only accessed by agent $i$; $\frac{\beta}{2m} \|\tilde{x}\|^2$ is a regularisation term used to avoid overfitting of the data. We set $m = 100$ and $\beta = 0.1$ in this simulation and utilise the data set from MNIST handwritten digit database [43]. A subset of 12,000 samples, 8 and 9 labelled as $-1$ and $+1$, is chosen from MNIST, from which $N = 8000$ random samples are chosen as training set and the rest of the samples are used for testing. Figure 7 shows a part of randomly selected training samples from 12,000 samples, where each image is featured as a 784-dimensional vector. Figure 8 depicts a fixed strongly connected directed network with $m = 100$ agents. The performance of the tested algorithms shown in Figure 9 demonstrates that the proposed algorithm shows accelerated linear convergence than existing well-known algorithms over a fixed directed network. Figure 10 depicts the testing error rate of the tested algorithms.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we concentrated on studying distributed optimisation over time-varying directed networks. After reviewing the distributed heavy-ball method, the distributed Nesterov method, and TV-AB algorithm, a double accelerated distributed optimisation algorithm was proposed. Then, based on the analysis of matrix inequalities, we have shown that the proposed algorithm converges linearly to the global optimal solution. Moreover, the proposed algorithm was applied into fixed directed networks and also achieves a linear convergence rate. We emphasise that the acceleration was not proved analytically and was observed via a series of simulations, which is
standard in related literature as the technical analysis and so far remains an open problem in the theoretical aspect. As future work, it would be of considerable interest to theoretically analyse the acceleration to establish an explicit accelerated convergence rate.

ACKNOWLEDGEMENTS

The work described in this paper was supported by the Fundamental Research Funds for the Central Universities (Grant No. XDJK2019AC001); by the Innovation Support Program for Chongqing Overseas Returnees (Grant No. cx2019005); by the National Natural Science Foundation of China (Grant Nos. 61773321 and 61762020); by the China Scholarship Council (Grant No. 20180852063); the Science and Technology Talents Fund for Excellent Young of Guizhou (Grant No. QKHJRC20195669); the Science and Technology Foundation of Guizhou (Grant No. QKHJC20181083); the Science and Technology Top-Notch Talents Support Project of Colleges and Universities in Guizhou (Grant No. QJHKY2016065); the High-Level Innovative Talents Project of Guizhou (Grant No. QRLF201621).

REFERENCES

1. Wai, H.T., et al.: Multi-agent reinforcement learning via double averaging primal-dual optimization. Advances in Neural Information Processing Systems, pp. 9649–9660. Curran Associates, New York (2018)
2. Raja, H., Bajwa, W.U.: Cloud K-SVD: A collaborative dictionary learning algorithm for big, distributed data. IEEE Trans. Signal Process. 64(1), 173–188 (2016)
3. Li, Z., et al.: A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. IEEE Trans. Signal Process. 67(17), 4944–4956 (2019)
4. Mohraz, A., et al.: A decentralized second-order method with exact linear convergence rate for consensus optimization. IEEE Trans. Signal Inf. Process. Netw. 2(4), 507–522 (2016)
5. Jadabaie, A., et al.: Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans. Autom. Control 48(6), 988–1001 (2003)
6. Chen, J., Sayed, A.H.: Diffusion adaptation strategies for distributed optimization and learning over networks. IEEE Trans. Signal Process. 60(8), 4289–4305 (2012)
7. Rabbat, M., Nowak, R.: Distributed optimization in sensor networks. In: Third International Symposium on Information Processing in Sensor Networks, pp. 20–27. IEEE, Piscataway (2004)
8. Li, H., et al.: Distributed robust algorithm for economic dispatch in smart grids over general unbalanced directed networks. IEEE Trans. Industr. Inform. 16(7), 4322–4332 (2020)
9. Li, K., et al.: Distributed optimisation based on multi-agent system for resource allocation with communication time-delay. IET Control Theory Appl. 14(4), 549–557 (2020)
10. Tsitsiklis, J.N., et al.: Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Trans. Autom. Control 31(9), 803–812 (1986)
11. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control 54(1), 48–61 (2009)
12. Shi, W., et al.: On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. 62(7), 1750–1761 (2014)
13. Terelius, H., et al.: Decentralized multi-agent optimization via dual decomposition. IFAC Proc. Vol. 44(1), 11245–1125 (2011)
14. Mota, J.F.C., et al.: D-ADMM: A communication-efficient distributed algorithm for separable optimization. IEEE Trans. Signal Process. 61(10), 2718–2723 (2013)
15. Jakovetic, D., et al.: Fast distributed gradient methods. IEEE Trans. Autom. Control 59(5), 1131–1146 (2014)
16. Shi, W., et al.: EXTRA: An exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. 25(2), 944–966 (2015)
17. Yuan, K., et al.: Exact diffusion for distributed optimization and learning - Part I: Algorithm development. IEEE Trans. Signal Process. 67(3), 708–723 (2019)
18. Yuan, K., et al.: Exact diffusion for distributed optimization and learning - Part II: Convergence analysis. IEEE Trans. Signal Process. 67(3), 724–739 (2019)
19. Nedić, A., Olshevsky, A.: Distributed optimization over time-varying directed graphs. IEEE Trans. Autom. Control 60(3), 601–615 (2015)
20. Tsianos, K.L., et al.: Push-sum distributed dual averaging for convex optimization. In: IEEE Conference on Decision and Control, pp. 5453–5458. IEEE, Piscataway (2013)
21. Xi, C., Khan, U.A.: DEXTRA: A fast algorithm for optimization over directed graphs. IEEE Trans. Autom. Control 62(10), 4980–4993 (2017)
APPENDIX A: PROOF OF LEMMA 4
Proof. Recalling \( \hat{x}_k^p = g_k - (1_m^T \otimes I_d)g_k \), we have
\[
(1_m^T \otimes I_d)g_k = (1_m^T \otimes I_d)VF(\hat{x}_k).
\]
(A1)

Then, due to the fact that \((1_m^T \otimes I_d)VF(x^*) = 0\), it holds
\[
\|g_k\| = \|\hat{x}_k^p + (1_m^T \otimes I_d)(VF(s_k) - VF(x^*))\| \\
\leq \|\hat{x}_k^p\| + L_f\|1_m\|\|s_k - x^*\|, \\
\leq \|\hat{x}_k^p\| + L_f\sqrt{m}\|1_m\|\|s_k - x^*\|, \\
\leq \|\hat{x}_k^p\| + mL_f\|s_k - x^*\|,
\]
(A2)

where we apply Lipschitz continuity of \(F\) in the first inequality and the second inequality uses the Jensen’s inequality. In next step, we continue to bound \(\|s_k - x^*\|\). From (11c), we have
\[
\|s_k - x^*\| = \|(I_m + D_\beta)\lambda_k - D_\beta A_{k-1}x_{k-1} - x^*\| \\
\leq \|\hat{x}_k^p\| + \|r_k\| + \tilde{\beta}\|d_k\|. \\
\]
(A3)

Combining (A2) with (A3) yields
\[
\|g_k\| \leq mL_f\|\hat{x}_k^p\| + mL_f\|r_k\| + \tilde{\beta}mL_f\|d_k\| + \|\hat{x}_k^p\|. 
\]

Noting that \(\|r_k\| \leq \|V_k \otimes I_d\|\|g_k\|\) and \(\|V_k \otimes I_d\| = \|V_k\| = \max_{i} |v_{ik}| < 1\), we therefore have
\[
\|r_k\| \leq mL_f\|\hat{x}_k^p\| + mL_f\|d_k\| + \tilde{\beta}mL_f\|d_k\| + \|\hat{x}_k^p\|. \\
\]
(A4)

□

APPENDIX B: PROOF OF LEMMA 5
Proof. According to (11b), it holds
\[
\{x_{k+1} = A_kx_k + D_\beta (A_{k+1} + I_{mn})(x_k - A_{k-1}x_{k-1}) - D_\beta y_k \}
\]
\[
= \sum_{l=0}^{\bar{c}+1} A_{k,l} (D_\beta (A_{k+1} + I_{mn})d_{k+1} - D_\beta y_k) + A_{k,\bar{c}}x_{k-\bar{c}+1},
\]
(B1)

which yields
\[
\|\hat{x}_{k+1}^p\| \\
= \|x_{k+1} - 1_m \otimes (\phi_{k+1} \otimes I_d) x_{k+1}\| \\
\leq \tilde{\beta}\| \left((I_m - 1_m \otimes \phi_{k+1} \otimes I_d) \sum_{l=0}^{\bar{c}+1} A_{k,l} (A_{k+1} + I_{mn})d_{k+1} \right) \| \\
\]

How to cite this article: Hu J, Yan Y, Li H, Wang Z, Xia D, Guo J. Convergence of an accelerated distributed optimisation algorithm over time-varying directed networks. IET Control Theory Appl. 2021;15:24–39. https://doi.org/10.1049/cth2.12022
\[ \text{APPENDIX C: PROOF OF LEMMA 6} \]

**Proof.** Notice that \( y_k = (\nu_k^T \otimes I_m) y_k + (V_k \otimes I_m) \tilde{z}_k \), it holds

\[
\| y_{k+1} \| = \| (1_m \phi_k^T \otimes I_m) x_{k+1} - 1_m \otimes \tilde{x}^k \| \\
\leq \| (1_m \phi_k^T \otimes I_m) x_{k} \| - D_k \phi_k^T v_k (1_m \nu_k \otimes I_m) \nabla F (y_k) - 1_m \otimes \tilde{x}^k \\
+ \tilde{\beta} \| (1_m \phi_k^T \otimes I_m) (x_k - A_k x_{k-1}) \| \\
+ \tilde{\alpha} \sqrt{m} \| z_k \| + 2 \tilde{\beta} \sqrt{m} \| x_k - A_k x_{k-1} \| \\
\leq \| (1_m \phi_k^T \otimes I_m) x_{k} \| - D_k \phi_k^T v_k (1_m \nu_k \otimes I_m) \nabla F (y_k) - 1_m \otimes \tilde{x}^k \\
+ \tilde{\alpha} \sqrt{m} \| z_k \| + 2 \tilde{\beta} \sqrt{m} \| x_k - A_k x_{k-1} \|,
\]

where the first inequality leverages (11b) and (11c). Then, from [11, Lemma 4], we know that \( \phi_k^T v_k \geq 1/m \) and \( \phi_k^T v_k \leq 1 \), for \( \forall k \geq 0 \). Thus, it holds

\[
\| (1_m \phi_k^T \otimes I_m) x_{k} \| - D_k \phi_k^T v_k (1_m \nu_k \otimes I_m) \nabla F (y_k) - 1_m \otimes \tilde{x}^k \\
\leq \sqrt{m} \| z_k \| - mD_k \phi_k^T v_k \nabla \tilde{f} (z_k) - \tilde{x}^k \\
+ \tilde{\alpha} \| 1_m \| \sum_{j=1}^{C-1} (\nabla f_j (z_k) - \nabla f_j (\tilde{x}_j)) \|
\]

\[
\text{APPENDIX D: PROOF OF LEMMA 7} \]

**Proof.** From (11b), one can verify

\[
\| r_{k+1} \| = \| A_k x_k + D_k (A_k + I_m) (x_k - A_{k-1} x_{k-1}) - D_k y_k - A_k x_k \| \\
\leq \tilde{\beta} \| A_k + I_m \| \| d_k \| + \tilde{\alpha} \| y_k \|.
\]

Substituting (A4) into (D1) reduces to

\[
\| d_{k+1} \| \leq \tilde{\alpha} m \| z_k \| + \tilde{\beta} \| A_k + I_m \| \| d_k \| + \tilde{\alpha} \| y_k \|
\]

\[
\text{APPENDIX E: PROOF OF LEMMA 8} \]

**Proof.** Define \( R_{k, \tilde{C}} = R_k \ldots R_{k-(\tilde{C}-1)} \) and recall (11d), we have

\[
\| z_{k+1} \| = R_{k+1} \| z_k \| + (V_{k+1}^{-1} \otimes I_m) (\nabla F (y_{k+1}) - \nabla F (y_k)) \\
= R_{k} (R_{k-1} \ldots R_{k-(\tilde{C}-1)}) (V_{k}^{-1} \otimes I_m) (\nabla F (y_{k}) - \nabla F (y_{k-1})) \\
+ (V_{k}^{-1} \otimes I_m) (\nabla F (y_{k}) - \nabla F (y_{k})).
\]
 Consequently, for $\forall k \geq 0$, taking Euclidean norm on both sides of (E1) yields

$$\begin{align*}
\|x_{k+1}^v\| &= \left\|x_{k+1} - (1 - \nu_{k+1}^T \otimes I) x_{k+1} \right\| \\
&\leq \left\| \left( I_m - (1 - \nu_{k+1}^T \otimes I) \right) R_{k,\bar{C},\bar{C}_k+1} \right\| \\
&+ \sum_{j=0}^{C-1} \left\| \left( I_m - (1 - \nu_{k+1}^T \otimes I) \right) R_{k,j} \left( V_{k-1}^{-1} - (j-1) \otimes I \right) \right\| J_j \left( \nu_{k-1} - \nu_{k-1} \right) \\
&\leq \gamma B \|x_{k,C+1}^v\| + m^{\mu_C} Q_h L_j \sum_{j=0}^{C-1} \|x_{k-1+j} - s_{k-j}\|. \tag{E2}
\end{align*}$$

where the last inequality utilises Corollary 2 and Jensen's inequality. To proceed, we bound $\|s_{k-1+j} - s_{k-j}\|$. From (11c), it holds

$$\begin{align*}
\|s_{k-1+j} - s_{k-j}\| &= \left\| \left( I_m + D_\beta \right) x_{k-1+j} \right\| \\
&- \left( (A_{k-j} + I_m) D_\beta + I_m \right) x_{k-j} + D_\beta A_{k-j+1} x_{k-j+1} \\
&\leq \beta \left( 1 + (1 + \beta) \left\| A_{k-j} + I_m \right\| + \alpha \left( 1 + \beta \right) m L_j \right) \|d_{k,j}\| \\
&+ \alpha \left( \beta + 1 \right) \left\| x_{k-j}^v \right\| + \alpha \left( \beta + 1 \right) m L_j \| \gamma \| \|x_{k-j}^v\|. \tag{E3}
\end{align*}$$

Substituting (E3) into (E2) reduces to

$$\begin{align*}
\|x_{k+1}^v\| &\leq \sum_{j=0}^{C-1} \beta \left( 1 + (1 + \beta) \left\| A_{k-j} + I_m \right\| + \alpha m L_j \left( 1 + \beta \right) \right) \|d_{k,j}\| \\
&+ \sum_{j=0}^{C-1} \left( \| A_{k-j} - I_m \| + \alpha m L_j \left( \beta + 1 \right) \right) \left\| x_{k-j}^v \right\| \\
&+ \sum_{j=0}^{C-1} \alpha m L_j \left( \beta + 1 \right) \| \gamma \| \|x_{k-j}^v\| \\
&+ \gamma B \|x_{k,C+1}^v\|. \tag{E4}
\end{align*}$$

where $t = \frac{1}{L_j m^{\mu_C} Q_h} > 0$. 

**APPENDIX F: DEFINITION OF PARAMETERS**

For the clarity of exposition, we place the definitions of parameters $\mathcal{W}_i, \omega_i$, and $\tilde{\omega}_i$ in this section.

$$\begin{align*}
\mathcal{W}_1 &= \omega_k^i + \tilde{\alpha} \tilde{\beta} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_2 &= \tilde{\beta} \omega_k^i + \tilde{\beta}^2 \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \tilde{\beta}^2 \omega_k^i, \\
\mathcal{W}_3 &= \omega_k^i + \tilde{\alpha} \tilde{\beta} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_4 &= \tilde{\beta} \omega_k^i + \tilde{\beta}^2 \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \tilde{\beta}^2 \omega_k^i, \\
\mathcal{W}_5 &= \tilde{\beta} \omega_k^i + \tilde{\alpha} \tilde{\beta} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_6 &= \tilde{\beta} \omega_k^i + \tilde{\beta}^2 \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \tilde{\beta}^2 \omega_k^i, \\
\mathcal{W}_7 &= \gamma B + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_8 &= \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_9 &= \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i, \\
\mathcal{W}_{10} &= \gamma B + \tilde{\alpha} \omega_k^i + \tilde{\alpha} \omega_k^i.
\end{align*}$$

**PARAMETERS**

\[ \begin{align*}
\omega_1 &= m L_f Q_A, \\
\omega_2 &= Q_A, \\
\omega_3 &= m L_f, \\
\omega_4 &= \frac{\mu}{m^{\mu_C} - 1}, \\
\omega_5 &= \sqrt{m}, \\
\omega_6 &= \|A_k + I_m\|, \\
\omega_7 &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_8 &= L_j m^{\mu_C} Q_B, \\
\omega_9 &= m L_f Q_B, \\
\omega_{10} &= \|A_k + I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{11} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{12} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{13} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{14} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{15} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{16} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{17} &= \|A_k + I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{18} &= \|A_k - I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{19} &= \|A_k + I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{20} &= m L_f Q_A, \\
\omega_{21} &= m L_f Q_B, \\
\omega_{22} &= \|A_k\| \|A + I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{23} &= \|A_k\| \|A + I_m\| L_j m^{\mu_C} Q_B, \\
\omega_{24} &= L_j \|B_{\alpha}\|.
\end{align*} \]
\[ \omega_{25} = \|A + I_{\text{mm}}\|, \]
\[ \omega_{26} = L_f \|I_{\text{mm}} - A\| / \|I_{\text{mm}} - B_\infty\|, \]
\[ \omega_{27} = L_f^2, \]
\[ \omega_{28} = L_f \|I_{\text{mm}} - B_\infty\|, \]
\[ \omega_{29} = L_f \|A + I_{\text{mm}}\| / \|I_{\text{mm}} - B_\infty\|, \]
\[ \hat{\omega}_1 = \omega_7 + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \omega_8, \]
\[ \hat{\omega}_2 = \bar{\beta} \omega_9 + \bar{\beta} \omega_{10} + \bar{\beta}^2 \omega_{10} + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \bar{\beta}^2 \omega_8, \]
\[ \hat{\omega}_3 = \omega_{11} + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \omega_8, \]
\[ \hat{\omega}_4 = \bar{\beta} \omega_9 + \bar{\beta} \omega_{12} + \bar{\beta}^2 \omega_{12} + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \bar{\beta}^2 \omega_8, \]
\[ \hat{\omega}_5 = \omega_{13} + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \omega_8, \]
\[ \hat{\omega}_6 = \bar{\beta} \omega_9 + \bar{\beta} \omega_{14} + \bar{\beta}^2 \omega_{14} + \bar{\alpha} \bar{\beta} \omega_8 + \bar{\alpha} \bar{\beta}^2 \omega_8, \]
\[ \hat{\omega}_7 = \gamma_B + \bar{\alpha} \bar{\beta} \omega_9 + \bar{\alpha} \omega_9, \]
\[ \hat{\omega}_8 = \sup_{k \geq 0} \|A_k + I_{\text{mm}}\|, \]
\[ \hat{\omega}_9 = \sup_{k \geq 0} \|A_k - I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{10} = \sup_{k \geq 0} \|A_k + I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{11} = \sup_{l \geq 0} \|A_l - I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{12} = \sup_{l \geq 0} \|A_l + I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{13} = \sup_{k \geq \tilde{C}^{-1}} \|A_{k-\tilde{C}^{-1}} - I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{14} = \sup_{k \geq \tilde{C}^{-1}} \|A_{k-\tilde{C}^{-1}} + I_{\text{mm}}\| L_f m^{\text{mm}} \mathcal{Q}_B, \]
\[ \hat{\omega}_{15} = \sup_{l \geq 0} \|A_l - I_{\text{mm}}\|, \]
\[ \hat{\omega}_{16} = \sup_{k \geq \tilde{C}^{-1}} \|A_{k-\tilde{C}^{-1}} + I_{\text{mm}}\|. \]

This is the end of the section.