2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Exact Noise-robust Distributed Gradient-tracking Algorithm for Constraint-Coupled Resource Allocation Problems

Wenwen Wu[1], Shanying Zhu[1], Shuai Liu[2] and Xinping Guan[1]

***Abstract*— The Industrial Internet of Things (IIoT) gradually becomes a new paradigm for information exchange in the industrial production environment. To ensure the high reliability of IIoT services, an efficient resource allocation method with good robustness is urgently needed under complex industrial environments. This paper considers the distributed constraint-coupled resource allocation problem with noisy information exchange over an undirected network, where each agent holds a private cost function and obtains the solution via only local communications. Communication noise poses a challenge to gradient-tracking based algorithm as the impact of noise will accumulate and its variance tends to infinity when the noise is persistent. Adopting noise-tracing scheme, we propose an exact noise-robust distributed gradient-tracking algorithm to achieve cost-optimal distribution of resources, which can avoid noise-accumulation in the tracking step. Moreover, noise suppression parameters are introduced to further attenuate the impact of noise. With diminishing suppression parameters, it is theoretically proved that the proposed algorithm is able to achieve exact convergence to the optimal solution. Finally, a numerical example is provided for verification.**

I. Introduction

As an emerging and prospective paradigm, the Industrial Internet of Things (IIoT) enables intelligent manufacturing via the interconnection and interaction of industrial production elements. Distributed constraint-coupled resource allocation problem (DRAP) is an important term associated with the IIoT, providing a efficient way to make systems more flexible and computation friendly [1], e.g. it has wide applications in software-defined networks [2] and MEC systems [3]. The goal of DRAP is achieving the cost-optimal distribution of limited resources among users to meet their demands, local constraints, and possibly certain coupled global constraints. In this paper, we focus on solving the following DRAP with a linear coupled constraint

$$\begin{aligned} &\min_{W\in\mathbb{R}^{np}} f(W)=\sum_{i=1}^{n} f_i(\mathbf{w}_i)\\ &s.t.\ \sum_{i=1}^{n} A_i\mathbf{w}_i=\sum_{i=1}^{n}\mathbf{d}_i,\ \ \mathbf{w}_i\in\Omega_i,\forall i, \end{aligned} \tag{1}$$

*This work was supported in part by National Key R&D Program of China under the grant 2022YFB3303900 and the NSF of China under the grants 62173225, 62103272 and the Key Program of the National Natural Science Foundation of China under Grant 62133008.

[1]Wenwen Wu, Shanying Zhu and Xinping Guan are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China; Key Laboratory of System Control and Information Processing, Ministry of Education of China,Shanghai 200240, China, and also Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai 200240, China.

[2]Shuai Liu is with the School of Control Science and Engineering, Shandong University, Jinan 250061,China.

where $f_i:\mathbb{R}^p\to\mathbb{R}$ is the agent i's private cost function, $W=\left[\mathbf{w}_1^T,\cdots,\mathbf{w}_n^T\right]^T$, $\mathbf{w}_i\in\mathbb{R}^p$ is the decision vector of agent i and $\mathbf{d}_i$ denotes the local resource demand. Ω_i is a local convex and compact set which encodes local constraints of agent i. $A_i\in\mathbb{R}^{m\times p}$ is the nonzero coupling matrix.

Compared with centralized methods, distributed ones that operate with only local information have better scalability and robustness to possible node failures, especially for large-scale systems [4]. There are extensive efforts devoted to solving optimization problems with only simple coupling between agents, i.e., all A_i are identity matrices, from either the primal or the dual perspectives [5]–[7]. Considering DRAP (1), distributed algorithms are proposed in [8], where the tracking technique is employed to track the constraint deviation. With Ω_i in (1) being a polyhedron, ref. [9] proposed a dual consensus ADMM method. And a proximal diffusion strategy is developed in [10] and its convergence is established even in the presence of nonsmooth terms.

It is noted that aforementioned distributed algorithms have been designed under the assumption of ideal communication networks without any distortion and noise. In practice, the communication channels might be corrupted by additive noise [11]. Quantization before transmission to reduce communication burden is another source of communication noise [12]–[14]. To alleviate the impact induced by quantization, the diminishing noise suppression parameter is introduced in [14]. Ref. [15] proposed a gradient-tracking based algorithm to solve optimization problems with stochastic gradient. To avoid the noise-accumulation problem issue, the noise-tracing scheme is proposed in [16] and it ensures that the global gradient estimation is unbiased with bounded variance. By incorporating diminishing suppression parameters with noise-tracing scheme, a distributed algorithm is proposed in [17], which has guaranteed optimality even under noisy interference. We note that [14]–[17] only deals with the special case of DRAP with only consensus constraints. Extension to the DRAP with a simple coupling constraint is performed in [18]. Moreover, a dual method is proposed in [19], but it can only reach a neighborhood of the optimum. As for solving the general DRAP (1), an nosie-robust algorithm is proposed in [20] based on the stochastic approximation technique. However, it requires exchanges of global optimization variables among all agents.

In this paper, a fully distributed optimization algorithm (ERDGA) is proposed to solve constraint-coupled DRAP (1) with noisy information exchange. Compared with works in [14]–[19], the proposed algorithm can tackle a more general coupling constraint between agents. As for noise

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383568

treatment, the noise-tracing scheme is introduced to obtain the unbiased estimate of the global gradient with bounded variance. To further eliminate steady state errors induced by persistent noise, two diminishing noise suppression parameters are implemented, which poses a challenge to conventional convergence analysis methods of gradient-tracking based algorithms in [16], [19], [21]. Using martingale theory, we theoretically prove that the ERDGA can achieve exact convergence to the optimal solution for strongly convex cost functions, which are not necessarily to be smooth.

The rest of the paper is organized as follows. Section II provides preliminaries. The proposed algorithm is developed in Section III and the theoretical analysis of its convergence is given in Section IV. Then the algorithm is numerically tested in Section V. Finally, Section VI concludes the paper.

Notations: Vectors default to columns if not otherwise specified. The Kronecker product is denoted by $\otimes$. Let $\mathbf{1}_n$ be the n-dimension vector with all one entries. For vectors, $\|\cdot\|$ denotes the 2-norm. For matrices, $\|\cdot\|$ denotes the spectral norm. $\preccurlyeq$ denotes the element-wise less than or equal to. We use $blkdiag(X_1,\cdots,X_n)$ to refer to the block-diagonal matrix with $X_1,\cdots,X_n$ as blocks. For a random variable x, we use $\mathbb{E}[x]$ to denote its expectation.

II. Preliminaries

We make the following assumptions on the DRAP (1):

Assumption 1: Each cost function $f_i(\mathbf{w})$ is strongly convex, i.e., for any $\mathbf{w},\mathbf{w}' \in \mathbb{R}^p$, the relation $(\mathbf{w}-\mathbf{w}')^T(\nabla f_i(\mathbf{w})-\nabla f_i(\mathbf{w}')) \geq \mu_i\|\mathbf{w}-\mathbf{w}'\|^2$ holds, where $\mu_i>0$. Define $\mu=\min_i\{\mu_i\}$.

The communication network over which agents exchange information can be represented by an undirected graph $\mathcal{G}=(\mathcal{N},\mathcal{E})$, where $\mathcal{N}=\{1,\cdots,n\}$ is the set of agents, $\mathcal{E}\subseteq\mathcal{N}\times\mathcal{N}$ denotes the set of edges, accompanied with a nonnegative weighted matrix $\mathcal{W}=[w_{ij}]$. For any $i,j\in\mathcal{N}$ in the network, $w_{ij}>0$ denotes agent j can exchange information with agent i. The collection of all individual agents that agent i can communicate with is defined as its neighbors set $\mathcal{N}_i$.

Assumption 2: The graph $\mathcal{G}$ is undirected and connected.

Under Assumption 2, the weight matrix $\mathcal{W}\in\mathbb{R}^{n\times n}$ is doubly stochastic, i.e., $\mathcal{W}\mathbf{1}_n=\mathbf{1}_n$ and $\mathbf{1}_n^T\mathcal{W}=\mathbf{1}_n^T$.

Assumption 3: The problem satisfies Slaters condition.

Assumptions 1-3 are standard when solving related problems. Under Assumptions 1, 3, the strong duality holds. Specially, Assumptions 3 is not needed when Ω_i is a polyhedron.

In this paper, we focus on the error caused by noisy communication links and/or quantization. In practice, the communication channels can be corrupted by additive noise, which was statistically modeled as Gaussian in [22]. In addition, to alleviate heavy load to communication networks, quantization techniques are particularly critical. Here, we take the following random quantization scheme [14] for example. For a single number $x\in[l,u]$, we uniformly divide the interval into B bins, whose end points are bounded by τ_i, i.e., $l=\tau_1\leq\cdots\leq\tau_B=u$ and $\triangle\triangleq\tau_{i+1}-\tau_i=\frac{u-l}{B-1}$. Thus, $b=log_2(B)$ bits can be used to index the $\{\tau_i\}$. Given $x\in[\tau_i,\tau_{i+1})$, we assign a probability based on its relative location inside this interval and choose either τ_i or τ_{i+1} to represent x at random: $Q(x)=\tau_i$, w.p. $1-\frac{x-\tau_i}{\triangle}$, and $Q(x)=\tau_{i+1}$, w.p. $\frac{x-\tau_i}{\triangle}$, where the random variable $Q(x)$ thus satisfies $\mathbb{E}[Q(x)]=x,\ \mathbb{E}[(Q(x)-x)^2]\leq\frac{\triangle^2}{4}$.

In the following parts, we use two independent random sequences $\{\mathbf{n}_{X_k}\}_{k>0}$, $\{\mathbf{n}_{L_k}\}_{k>0}$ to summarize the aforementioned communication noise: $\mathbf{n}_{X_k}:=[\mathbf{n}_{\mathbf{x}_{1,k}}^T,\cdots,\mathbf{n}_{\mathbf{x}_{n,k}}^T]^T$, $\mathbf{n}_{L_k}:=[\mathbf{n}_{\mathbf{l}_{1,k}}^T,\cdots,\mathbf{n}_{\mathbf{l}_{n,k}}^T]^T$, where $\mathbf{n}_{\mathbf{x}_{i,k}},\mathbf{n}_{\mathbf{l}_{i,k}}\in\mathbb{R}^m$ denote the noise encountered by agent i at iteration time k.

Assumption 4: Noises $\mathbf{n}_{X_k}$, $\mathbf{n}_{L_k},\forall k>0$ have zero mean and bounded variance, i.e., $\mathbb{E}[\mathbf{n}_{X_k}]=\mathbb{E}[\mathbf{n}_{L_k}]=0$ and $\mathbb{E}[\|\mathbf{n}_{X_k}\|^2]\leq\sigma_X^2$, $\mathbb{E}[\|\mathbf{n}_{L_k}\|^2]\leq\sigma_L^2$ for some σ_X, $\sigma_L>0$.

III. Algorithm Development

In this section, we will develop an exact noise-robust distributed algorithm based on Lagrangian duality.

Introducing Lagrange multiplier $\mathbf{x}$, we construct the Lagrangian $L(W,\mathbf{x})$. Then, the dual problem of DRAP (1) can be derived, which is equivalent to the following distributed consensus problem (e.g., see [6] for details)

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^m}\quad & \sum_{i=1}^n F_i(\mathbf{x})=\sum_{i=1}^n F_i(\mathbf{x}_i)\\ s.t.\quad & \mathbf{x}_1=\mathbf{x}_2=\cdots=\mathbf{x}_n=\mathbf{x}\in\mathbb{R}^m,\end{aligned}\tag{2}$$

in which the local cost function is $F_i(\mathbf{x})\triangleq f_i^*\left(-A_i^T\mathbf{x}\right)+\mathbf{x}^T\mathbf{d}_i$, where $f_i^*\left(-A_i^T\mathbf{x}\right)$ is the convex conjugate function $f_i^*\left(-A_i^T\mathbf{x}\right)=\sup_{\mathbf{w}\in\Omega_i}\left(-\mathbf{x}^TA_i\mathbf{w}-f_i(\mathbf{w})\right)$.

Under Assumption 1, f^* is differentiable, $\frac{1}{\mu}$-Lipschitz smooth, and the supremum related to f^* is attainable [23]. It follows from Proposition B.25 in [24] that the gradient of $F_i(\mathbf{x})$ can be obtained as $\nabla F_i(\mathbf{x})=-A_i\nabla f_i^*\left(-A_i^T\mathbf{x}\right)+\mathbf{d}_i=-A_i\cdot\arg\min_{\mathbf{w}\in\Omega_i}\left\{f_i(\mathbf{w})+\mathbf{x}^T\cdot A_i\mathbf{w}\right\}+\mathbf{d}_i$.

Adopting gradient-tracking scheme in the above dual problems (2) shows effectiveness in solving problem (1) in noiseless situations [25]. However, with noisy information exchange, such scheme suffers from poor convergence properties due to noise-accumulation:

$$\sum_{i=1}^n\mathbf{l}_{i,k+1}=\sum_{i=1}^n\left(\nabla F_i(\mathbf{x}_{i,k+1})+\sum_{t=0}^{k-1}\mathbf{n}_{\mathbf{l}_{i,t}}\right),\tag{3}$$

where the variable $\mathbf{l}_{i,k+1}$ is used to directly track the gradient information and its variance tends to infinity when the noise is persistent, making it unreliable. To avoid noise-accumulation, motivated by [16], [19], we adopt the variable $\mathbf{l}_{i,k+1}$ to record the impact of noise on gradient information at iteration k instead, which is named as noise-tracing

$$\mathbf{l}_{i,k+1}=\sum_{j=1}^n w_{ij}\cdot\mathbf{l}_{j,k}+\nabla F_i(\mathbf{x}_{i,k+1})+\mathbf{n}_{\mathbf{l}_{i,k}},\forall i\in\mathcal{N}.\tag{4}$$

Here, $\mathbf{l}_{i,k+1}-\mathbf{l}_{i,k}$ is used for gradient-tracking at iteration k, where the effect of the noise at iteration $k-1$ can thus be eliminated. When the initial condition $\sum_{i=1}^n\mathbf{l}_{i,0}=\sum_{i=1}^n\nabla F_i(\mathbf{x}_{i,0})$ holds, it can be derived by induction that

$$\sum_{i=1}^n\left(\mathbf{l}_{i,k+1}-\mathbf{l}_{i,k}\right)=\sum_{i=1}^n\left(\nabla F_i(\mathbf{x}_{i,k+1})+\mathbf{n}_{\mathbf{l}_{i,k}}\right),\tag{5}$$

i.e., the tracked information is the unbiased estimate of the global gradient with bounded variance under Assumption 4.

Additionally, two noise suppression parameters η_k, γ_k are introduced to alleviate the impact of noise. These two parameters determine the degree to which variables from the neighbors should be weighed against the local one when proceeding algorithm updating. As shown in Algorithm 1 below, the introduction of η_k, γ_k causes the noise that actually affects algorithm iterations to be $\eta_k \cdot \mathbf{n}_{\mathbf{x}_{i,k}}$, $\gamma_k \cdot \mathbf{n}_{\mathbf{l}_{i,k}}$.

Intuitively, without noise-accumulation, the error induced by noise tends to zero as two noise suppression parameters decay to 0. However, since η_k, γ_k also closely related to the consensus process, the consensus constraint in (2) may be difficult to satisfy if simply reduce two parameters to zeros with iteration. Thus, a proper design of η_k, γ_k is further needed, which will be elaborated in the next section.

We summarize the **E**xact noise-**r**obust **d**istributed **g**radient-tracking **a**lgorithm (ERDGA) in Algorithm 1.

Algorithm 1 : ERDGA algorithm

1: Parameters: $\mathcal{W} = [w_{ij}]$, $\eta_k, \gamma_k > 0$, $\beta_k > 0$, $\forall k$;
Initialization: Arbitrary $\mathbf{x}_{i,0} \in \mathbb{R}^m, \mathbf{w}_{i,0} \in \mathbb{R}^p, \mathbf{l}_{i,0} = -A_i\mathbf{w}_{i,0} + \mathbf{d}_i$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Dual variable update:
$$\mathbf{x}_{i,k+1} = (1-\eta_k)\,\mathbf{x}_{i,k} + \eta_k \left(\sum_{j=1}^n w_{ij} \cdot \mathbf{x}_{j,k} + \mathbf{n}_{\mathbf{x}_{i,k}}\right) - (\mathbf{l}_{i,k} - \mathbf{l}_{i,k-1}),$$
4: Primal variable update:
$$\mathbf{w}_{i,k+1} = \arg\min_{\mathbf{w}\in\Omega_i} \left\{ f_i(\mathbf{w}) + \mathbf{x}_{i,k+1}^T \cdot A_i\mathbf{w} \right\},$$
5: Auxiliary variable update:
$$\mathbf{l}_{i,k+1} = (1-\gamma_k)\,\mathbf{l}_{i,k} + \gamma_k \left(\sum_{j=1}^n w_{ij} \cdot \mathbf{l}_{j,k} + \mathbf{n}_{\mathbf{l}_{i,k}}\right) + \beta_k \left(-A_i\mathbf{w}_{i,k+1} + \mathbf{d}_i\right).$$
6: **end for**

Similar treatment can be found in [17]. Although problem (2) possesses similar structure with that in [17], we note that the convergence cannot be guaranteed by directly adopting their algorithm to solve (2), since an additional bounded assumption for dual variable is needed. This assumption is hard to satisfy when solving the DRAP (1) in this paper.

IV. Convergence Analysis of ERDGA

In this section, we will establish the convergence properties of the proposed algorithm.

For analysis, we define $X_k = [\mathbf{x}_{1,k}^T, \cdots, \mathbf{x}_{n,k}^T]^T$, $L_k = [\mathbf{l}_{1,k}^T, \cdots, \mathbf{l}_{n,k}^T]^T$, $\hat{\mathbf{1}} = \mathbf{1}_n \otimes I_m$, $\bar{\mathbf{x}}_k = \frac{\hat{\mathbf{1}}^T}{n} X_k$, $\bar{X}_k = \hat{\mathbf{1}}\bar{\mathbf{x}}_k$, $\bar{L}_k = \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n} L_k$, $F(X_k) = \sum_{i=1}^n F_i(\mathbf{x}_{i,k})$, $\nabla F(X_k) = [\nabla F_1(\mathbf{x}_{1,k})^T, \cdots, \nabla F_n(\mathbf{x}_{n,k})^T]^T$, $\boldsymbol{\mathcal{W}} = \mathcal{W} \otimes I_m$, $\mathbf{A} = blkdiag(A_1, \cdots, A_n)$. Then, we will use a deterministic counterpart of the supermartingale convergence result.

Lemma 1: [26] Let $\{U_k\}$, $\{V_k\}$ be non-negative vector sequences and $\{g_k\}$, $\{h_k\}$ be non-negative scalar sequences such that $U_{k+1} \preccurlyeq (P_k + g_k\mathbf{1}\mathbf{1}^T)U_k + h_k\mathbf{1} - Q_kV_k, \forall k > 0$ holds, where $\{P_k\}$, $\{Q_k\}$ are non-negative matrices. If $\{g_k\}$, $\{h_k\}$ satisfy $\sum_{k=1}^\infty g_k < \infty$ and $\sum_{k=1}^\infty h_k < \infty$, and there exists positive vector π such that $\pi^T P_k \le \pi^T$, $\pi^T Q_k \ge 0, \forall k > 0$ holds. Then we have 1) $\{\pi^T U_k\}$ is convergent; 2) $\{U_k\}$ is bounded; 3) $\sum_{k=1}^\infty \pi^T Q_k V_k < \infty$.

Denote $\boldsymbol{\mathcal{W}}_{\eta_k} := (1-\eta_k)I + \eta_k\boldsymbol{\mathcal{W}}$, $\boldsymbol{\mathcal{W}}_{\gamma_k} := (1-\gamma_k)I + \gamma_k\boldsymbol{\mathcal{W}}, \forall k > 0$. Under Assumption 2, the matrix $\boldsymbol{\mathcal{W}}_{\eta_k}$ has a unique right eigenvector $\hat{\mathbf{1}}$ (associated with eigenvalue 1) satisfying $\hat{\mathbf{1}}^T\hat{\mathbf{1}} = nI_m$ and so does the matrix $\boldsymbol{\mathcal{W}}_{\gamma_k}$. Moreover, the spectral radius of $\boldsymbol{\mathcal{W}}_{\eta_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n} = I + \eta_k(\boldsymbol{\mathcal{W}} - I) - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}$ is equal to $1 - \eta_k|\rho_{\boldsymbol{\mathcal{W}}-I}|$, where $\rho_{\boldsymbol{\mathcal{W}}-I}$ is an eigenvalue of $\boldsymbol{\mathcal{W}} - I$. Thus, we always have $\|\boldsymbol{\mathcal{W}}_{\eta_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}\| = 1 - \eta_k|\rho_{\boldsymbol{\mathcal{W}}-I}| < 1$. Similarly, we can derive that $\|\boldsymbol{\mathcal{W}}_{\gamma_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}\| = 1 - \gamma_k|\rho_{\boldsymbol{\mathcal{W}}-I}| < 1$. We use $\lambda_{\mathcal{W}} := |\rho_{\boldsymbol{\mathcal{W}}-I}|$ in the following parts for simplicity.

Lemma 2: Supposing that Assumptions 1-4 hold, the sequences generated by ERDGA satisfy:

$$m_{k+1,1} \le m_{k,1} + c_k^{1,2} m_{k,2} + \left(b_k^{1,3} + c_k^{1,3}\right) m_{k,3} + c_k^{1,4} m_{k,4} + d_k^1 - \mu\beta_{k-1}\mathbb{E}[\|W_k - W^\star\|^2]. \quad (6)$$

where $m_{k,1} = \mathbb{E}[\|\bar{X}_k - X^\star\|^2]$, $m_{k,2} = \mathbb{E}[F(\bar{X}_k)] - F(X^\star)$, $m_{k,3} = \mathbb{E}[\|X_k - \bar{X}_k\|^2]$, $m_{k,4} = \mathbb{E}[\|L_{k-1} - \bar{L}_{k-1}\|^2]$, $W^\star$, $X^\star$ are optimal primal and dual variables, respectively and parameters $b_k^{1,3}$, $c_k^{1,2}$, $c_k^{1,3}$, $c_k^{1,4}$, d_k^1 are given in appendix II.

Proof: See Appendix I.

The relation (6) shows that the iteration of $m_{k,1}$ is coupled with $m_{k,2}$, $m_{k,3}$ and $m_{k,4}$. For analysis, we use $M_k = [m_{k,1}, m_{k,2}, m_{k,3}, m_{k,4}]^T$ as a measure of convergence, where the first two terms quantify the optimality gap, and the rest terms quantify the consensus error among agents.

Lemma 3: Supposing that Assumptions 1-4 hold, the sequences generated by ERDGA satisfy:

$$M_{k+1} \preccurlyeq (B_k + C_k)M_k + D_k - H_kN_k, \quad (7)$$

where $N_k = [\mathbb{E}[\|W_k - W^\star\|^2], \mathbb{E}[\|\hat{\mathbf{1}}^T\nabla F(\bar{X}_k)\|^2]]^T$ and all matrices have nonnegative elements as shown in Appendix.

Proof: See Appendix II.

On the basis of Lemmas 1-3, the convergence of ERDGA is established in the following Theorem.

Theorem 1: Suppose that Assumptions 1-4 hold and positive sequences $\{\eta_k\}$, $\{\gamma_k\}$, and $\{\beta_k\}$ satisfy $\sum_{k=1}^\infty \eta_k = \infty$, $\sum_{k=1}^\infty \gamma_k = \infty$, $\sum_{k=1}^\infty \beta_k = \infty$, $\sum_{k=1}^\infty {\eta_k}^2 < \infty$, $\sum_{k=1}^\infty {\gamma_k}^2 < \infty$, $\sum_{k=1}^\infty \frac{{\beta_{k-1}}^2}{\eta_k} < \infty$, $\lim_{k\to\infty} \frac{\beta_{k-1}}{\eta_k} = 0$, $\lim_{k\to\infty} \frac{\gamma_{k-1}}{\eta_k} = \Phi$ for some bounded $\Phi > 0$. Then, the sequences generated by ERDGA satisfy (i) $\lim_{k\to\infty}\mathbb{E}[\|W_k - W^\star\|^2] = 0$, (ii) $\lim_{k\to\infty}\mathbb{E}[\|X_k - \bar{X}_k\|^2] = \lim_{k\to\infty}\mathbb{E}[\|L_k - \bar{L}_k\|^2] = 0$.

Proof: From Lemmas 2-3, we can obtain that

$$M_{k+1} \preccurlyeq (B_k + c_k\mathbf{1}\mathbf{1}^T)M_k + d_k\mathbf{1} - H_kN_k, \quad (8)$$

where c_k, d_k are equal to the max elements of C_k, D_k, respectively. Under the assumptions on sequences $\{\eta_k\}$, $\{\gamma_k\}$, $\{\beta_k\}$ in Theorem 1, we have $\sum_{k=1}^\infty c_k < \infty$, $\sum_{k=1}^\infty d_k < \infty$.

Define $\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \pi_4]^T$. We will prove that there exists a positive vector $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}^T B_k \le \boldsymbol{\pi}^T$ and $\boldsymbol{\pi}^T H_k \ge 0$ hold for $k \ge \bar{T}$ with some large enough $\bar{T} \ge 0$. It can be verified that the above two conditions hold when relations $\pi_1, \pi_2 \ge 0$, $\pi_3 \ge \frac{2\mu\|\mathbf{A}\|^2\pi_1 + \|\mathbf{A}\|^4\pi_2}{2\lambda_{\mathcal{W}}} \frac{\beta_{k-1}}{\eta_k}$, $\pi_4 \ge \frac{2\|\boldsymbol{\mathcal{W}}-I\|^2}{{\lambda_{\mathcal{W}}}^2} \frac{\gamma_{k-1}}{\eta_k}\pi_3$ are satisfied. The first two conditions

$\pi_1, \pi_2 \geq 0$ are easy to meet. Due to $\lim_{k\to\infty} \frac{\beta_{k-1}}{\eta_k} = 0$, for any given π_1, π_2, we can always find a bounded $\pi_3 > 0$ such that the third inequality holds. Moreover, a bounded $\pi_4 > 0$ can be found such that the forth inequality also holds after other elements of $\boldsymbol{\pi}$ are fixed since $\lim_{k\to\infty} \frac{\gamma_{k-1}}{\eta_k} = \Phi$. Thus, we can always find a $\boldsymbol{\pi}$ satisfying the above relations.

From Lemma 1, we have 1) $\{\pi^T M_k\}$ is convergent; 2) $\{M_k\}$ is bounded; 3) $\sum_{k=1}^{\infty} \pi^T H_k N_k < \infty$. So, the relation $\liminf_{k\to\infty} \mathbb{E}[\|W_k - W^\star\|^2] = 0$ holds since $\sum_{k=1}^{\infty} \beta_k = \infty$.

Since $\mathbf{w}_k \in \Omega_i$, the variable $\{W_k\}_{k>0}$ is bounded. We can find a convergent sub-sequence of $\{W_{k_i}\}_{k_i>0}$. Taking the limit along $k_i \to \infty$ yields $\lim_{k_i\to\infty} \mathbb{E}[\|W_{k_i} - W^\star\|^2] = 0$. Moreover, since W is a continuous function of X, $\{\boldsymbol{\pi}^T M_k\}$ is convergent and contractive with respect to the optimum, then we can derive that $\lim_{k\to\infty} \mathbb{E}[\|W_k - W^\star\|^2] = 0$ [27], i.e., the result (i) is proved.

Define $M'_k = \left[\mathbb{E}[\|X_k - \bar{X}_k\|^2], \mathbb{E}[\|L_{k-1} - \bar{L}_{k-1}\|^2]\right]^T$. From relation (7), we have

$$\begin{aligned} M'_{k+1} &\preccurlyeq (B'_k + C'_k) M'_k + D'_k \\ &\preccurlyeq (B'_k + c'_k \mathbf{1}\mathbf{1}^T) M'_k + d'_k \mathbf{1}, \end{aligned} \tag{9}$$

where c'_k, d'_k are equal to the max elements of C'_k, D'_k, respectively, and all matrices have nonnegative elements:

$$B'_k = \begin{bmatrix} b_k^{3,3} & b_k^{3,4} \\ 0 & b_k^{4,4} \end{bmatrix}, C'_k = \begin{bmatrix} c_k^{3,3} & 0 \\ c_k^{4,3} & 0 \end{bmatrix}, D'_k = \begin{bmatrix} c_k^{3,2} m_{k,2} + d_k^3 \\ c_k^{4,2} m_{k,2} + d_k^4 \end{bmatrix},$$

where parameters are given in appendix II. Note that $m_{k,2}$ is bounded as $\{M_k\}$ is proved to be bounded before. Under the assumptions on sequences $\{\eta_k\}$, $\{\gamma_k\}$, and $\{\beta_k\}$ in Theorem 1, we have $\sum_{k=1}^{\infty} c'_k < \infty$, $\sum_{k=1}^{\infty} d'_k < \infty$.

To prove (ii), we define $\boldsymbol{\pi}' = [\pi'_1, \pi'_2, \pi'_3, \pi'_4]^T$. Similarly, relations $\boldsymbol{\pi}'^T B'_k \leq (1 - \alpha\gamma_{k-1})\boldsymbol{\pi}'^T$ and $\boldsymbol{\pi}'^T H_k \geq 0$ hold when there exists a positive vector $\boldsymbol{\pi}'$ satisfying $(1 - \eta_k \lambda_{\mathcal{W}})\pi'_1 \leq (1 - \alpha\gamma_{k-1})\pi'_1$ and $\frac{\|\boldsymbol{\mathcal{W}} - I\|^2}{\lambda_{\mathcal{W}}} \frac{\gamma_{k-1}^2}{\eta_k} \pi'_1 + (1 - \gamma_{k-1}\lambda_{\mathcal{W}})\pi'_2 \leq (1 - \alpha\gamma_{k-1})\pi'_2$ with some constant $\alpha > 0$. It can be verified that the needed two relations hold when α, $\boldsymbol{\pi}'$ satisfy $\alpha \leq \min\{\lambda_{\mathcal{W}} \frac{\eta_k}{\gamma_{k-1}}, \lambda_{\mathcal{W}} - \frac{\pi'_1}{\pi'_2} \frac{\gamma_{k-1}}{\eta_k}\}$, and we can always find $\boldsymbol{\pi}'$ such that $\lambda_{\mathcal{W}} - \frac{\pi'_1}{\pi'_2} \frac{\gamma_{k-1}}{\eta_k}$ is greater than 0, since $\lim_{k\to\infty} \frac{\gamma_{k-1}}{\eta_k} = \Phi$ for some bounded $\Phi > 0$. By properly choosing vector $\boldsymbol{\pi}'$, we can always find such α for $k \geq \bar{T}$ for some large enough $\bar{T} \geq 0$. Thus, multiplying both sides of (9) by $\boldsymbol{\pi}'$, we have

$$\begin{aligned} \boldsymbol{\pi}'^T M'_{k+1} \leq & (1 + c'_k \frac{\boldsymbol{\pi}'^T \mathbf{1}}{\pi'_{\min}}) \boldsymbol{\pi}'^T M'_k + d'_k \boldsymbol{\pi}'^T \mathbf{1} \\ & - \alpha\gamma_{k-1} \boldsymbol{\pi}'^T M'_k, \end{aligned} \tag{10}$$

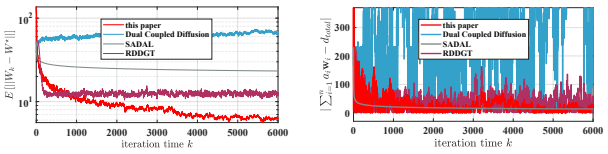
where the relation $\mathbf{1}^T \preccurlyeq \frac{\boldsymbol{\pi}'^T}{\pi'_{\min}}$ is used and $\pi'_{\min}$ is the minimum element of $\boldsymbol{\pi}'$. Similarly, from Lemma 1, we have that $\{\boldsymbol{\pi}'^T M'_k\}$ is convergent and $\sum_{k=1}^{\infty} \alpha\gamma_{k-1} \boldsymbol{\pi}'^T M'_k < \infty$. Thus, M'_k is convergent and $\lim_{k\to\infty} M'_k = 0$ holds, i.e., relation (ii) is proved. The proof is completed. ■

We emphasize that the requirement on the stepsize β_k and noise suppression parameters η_k, γ_{k-1} in Theorem 1 can be easily satisfied. For example, we can set $\beta_k = \mathcal{O}(k^{-a})$, $\eta_k = \gamma_{k-1} = \mathcal{O}(k^{-b})$ with $a, b \in (0.5, 1]$ and $2a - b > 1$.

V. Numerical Experiments

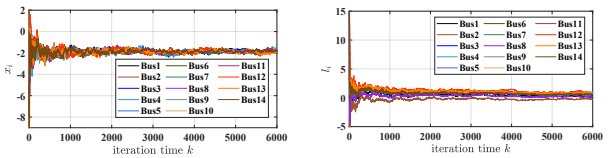
In this section, we conduct numerical experiments to verify our theoretical analysis. The proposed algorithm is tested by the economic dispatch of multi-microgrid systems with 14 microgrids [8]. We consider the problem with quadratic cost functions $f_i(\mathbf{w}_i) = t_i \mathbf{w}_i^2 + u_i \mathbf{w}_i + v_i, \forall i$. The coupling coefficients a_i are randomly chosen and the parameters of the generators are adopted from [19]. Set total load demand $d_{total} = \sum_{i=1}^{n} d_i = 231MW$. To model the noise interference, the exchanged information is corrupted by independent Gaussian noise $\mathcal{N}(0, 1)$. We set stepsize $\beta_k = \frac{0.1}{1+0.5\cdot k}$, suppression parameters $\eta_k = \gamma_k = \frac{1}{1+0.1\cdot k^{0.8}}$, which satisfy the conditions in Theorem 1.

We compare ERDGA with three state-of-the-art algorithms, termed Dual coupled diffusion algorithm [10], SADAL [20] and RDDGT [19], in terms of the error $\mathbb{E}\left[\|W_k - W^\star\|\right]$ and the expectation is approximated by averaging over 100 simulation results. The stepsize for Dual coupled diffusion algorithms is chosen as $\beta = 0.05$ and the stepsize for SADAL is the same as ERDGA. We set $\beta = 0.001, \eta = 0.08, \gamma = 0.5$ for RDDGT. The evolution of the errors is shown in Fig. 1 (a), where the error of RDDGT converges to a neighborhood of the optimum, while the Dual coupled diffusion algorithm has increasing errors instead. The errors of SADAL and ERDGA both continuously decrease. However, with same stepsize, our proposed algorithm has a faster convergence rate than SADAL. The evolution of the constraint violation $|\sum_{i=1}^{n} a_i \mathbf{w}_i - d_{total}|$ of the global coupled constraint is shown in Fig. 1 (b). It can be seen that the RDDGT has a low upper bound, while the Dual coupled diffusion algorithm can hardly satisfy the global coupled constraint. As for SADAL and ERDGA, they can gradually drive the constraint violation toward zero. Thus, the proposed ERDGA algorithm is shown to have better robustness to noise interference. Moreover, Fig. 2 (a) and Fig. 2 (b) show the evolution of dual variables $\mathbf{x}_i$, auxiliary variables $\mathbf{l}_i$, respectively. All dual variables gradually converge to the same value and so do the auxiliary variables, which is consistent with Theorem 1.



(a) (b)

Fig. 1. (a) Convergence error, and (b) constraint violation versus iteration time for different algorithms.



(a) (b)

Fig. 2. (a) Dual variables, and (b) auxiliary variables versus iteration time.

VI. Conclusion

In this paper, an exact noise-robust distributed gradient-tracking algorithm has been proposed to solve constraint-coupled resource allocation problems with noisy information exchange. By adopting the noise-tracing scheme and diminishing noise suppression parameters, the proposed algorithm has been shown to have better robustness to noise interference than existing distributed algorithms. Moreover, its exact convergence property to the optimum has been established for strongly convex cost functions. Finally, the theoretical results have been examined by numerical experiments.

References

[1] P. Goswami, A. Mukherjee, M. Maiti, S. K. S. Tyagi, and L. Yang, "A neural-network-based optimal resource allocation method for secure IIoT network," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2538–2544, 2022.

[2] H. Cao, J. Du, H. Zhao, D. X. Luo, N. Kumar, L. Yang, and F. R. Yu, "Toward tailored resource allocation of slices in 6G networks with softwarization and virtualization," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6623–6637, 2022.

[3] L. Tan, Z. Kuang, J. Gao, and L. Zhao, "Energy-efficient collaborative multi-access edge computing via deep reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 6, pp. 7689–7699, 2023.

[4] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[5] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "A dual splitting approach for distributed resource allocation with regularization," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 1, pp. 403–414, 2019.

[6] J. Zhang, K. You, and K. Cai, "Distributed dual gradient tracking for resource allocation in unbalanced networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2186–2198, 2020.

[7] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[8] W. Wu, S. Zhu, S. Liu, and X. Guan, "Differentially private distributed mismatch tracking algorithm for constraint-coupled resource allocation problems," in *Proc. 61st IEEE Conference on Decision and Control*, Cancún, Mexico, Dec. 2022, pp. 3965–3970.

[9] T.-H. Chang, "A proximal dual consensus ADMM method for multi-agent constrained optimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3719–3734, 2016.

[10] S. A. Alghunaim, K. Yuan, and A. H. Sayed, "A proximal diffusion strategy for multiagent optimization with sparse affine constraints," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4554–4567, 2020.

[11] N. M. Dehkordi, H. R. Baghaee, N. Sadati, and J. M. Guerrero, "Distributed noise-resilient secondary voltage and frequency control for islanded microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3780–3790, 2019.

[12] S. Liu, T. Li, L. Xie, M. Fu, and J.-F. Zhang, "Continuous-time and sampled-data-based average consensus with logarithmic quantizers," *Automatica*, vol. 49, no. 11, pp. 3329–3336, 2013.

[13] S. Zhu, C. Chen, J. Xu, X. Guan, L. Xie, and K. H. Johansson, "Mitigating quantization effects on distributed sensor fusion: A least squares approach," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3459–3474, 2018.

[14] T. T. Doan, S. T. Maguluri, and J. Romberg, "Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach," *IEEE Transactions on Automatic Control*, vol. 66, no. 10, pp. 4469–4484, 2021.

[15] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, vol. 187, pp. 409–457, 2021.

[16] S. Pu, "A robust gradient tracking method for distributed optimization over directed networks," in *Proc. 59th IEEE Conference on Decision and Control*, Jeju Island, Korea, Dec. 2020, pp. 2335–2341.

[17] Y. Wang and T. Başar, "Gradient-tracking-based distributed optimization with guaranteed optimality under noisy information sharing," *IEEE Transactions on Automatic Control*, vol. 68, no. 8, pp. 4796–4811, 2023.

[18] P. Yi, J. Lei, and Y. Hong, "Distributed resource allocation over random networks based on stochastic approximation," *Systems & Control Letters*, vol. 114, pp. 44–51, 2018.

[19] W. Wu, S. Liu, and S. Zhu, "Distributed dual gradient tracking for economic dispatch in power systems with noisy information," *Electric Power Systems Research*, vol. 211, p. 108298, 2022.

[20] N. Chatzipanagiotis and M. M. Zavlanos, "A distributed algorithm for convex constrained optimization under noise," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2496–2511, 2016.

[21] M. I. Qureshi and U. A. Khan, "Distributed saddle point problems for strongly concave-convex functions," *arXiv preprint arXiv:2202.05812*, 2022.

[22] J. G. Proakis, *Digital Communications*. New York, NY, USA: McGraw-Hill, 1995.

[23] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, vol. 306.

[24] D. Bertsekas, *Nonlinear Programming*. Belmont, Massachusetts: Athena Scientific, 2016.

[25] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *Proc. 54th IEEE Conference on Decision and Control*, Osaka, Japan, Dec. 2015, pp. 2055–2060.

[26] Y. Wang and A. Nedić, "Tailoring gradient methods for differentially-private distributed optimization," *IEEE Transactions on Automatic Control*, 2023, 10.1109/TAC.2023.3272968, early access.

[27] H. H. Bauschke and Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA, 2011.

Appendix I
Proof of Lemma 2

The proposed algorithm can be written as

$$X_{k+1} = \boldsymbol{\mathcal{W}}_{\eta_k} X_k - (L_k - L_{k-1}) + \eta_k \mathbf{n}_{X_k}, \tag{11}$$

$$L_{k+1} = \boldsymbol{\mathcal{W}}_{\gamma_k} L_k + \beta_k \nabla F(X_{k+1}) + \gamma_k \mathbf{n}_{L_k}, \tag{12}$$

where we use $\nabla F_i(\mathbf{x}_{i,k}) = -A_i \mathbf{w}_{i,k} + \mathbf{d}_i$. Under Assumption 1, $f_i^*(\mathbf{x})$ is convex, differentiable and it satisfies $\frac{1}{\mu}$-Lipschitz smoothness [23]. Thus, we have that $F(X)$ is convex and L'-Lipschitz smooth, where $L' := \frac{\|\mathbf{A}\|^2}{\mu}$.

It follows from the optimal condition of $\mathbf{w}$-update that $(\nabla f_i(\mathbf{w}_i^\star) + A_i \mathbf{x}_i^\star)^T (\mathbf{w}_{i,k} - \mathbf{w}_i^\star) \geq 0$, $(\nabla f_i(\mathbf{w}_{i,k}) + A_i \mathbf{x}_{i,k})^T (\mathbf{w}_i^\star - \mathbf{w}_{i,k}) \geq 0$, where $\mathbf{w}_i^\star$, $\mathbf{x}_i^\star$, are optimal primal, dual variables and thus relations $X^\star = \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n} X^\star$ and $\hat{\mathbf{1}}^T \nabla F(X^\star) = 0$ hold. Combining these two optimal conditions and summing it from $i = 1$ to n, we have

$$(X^\star - X_k)^T (\mathbf{A} W_k - \mathbf{A} W^\star) \geq \mu \|W_k - W^\star\|^2, \tag{13}$$

where the inequality follows from Assumption 1.

It follows from $\nabla F_i(\mathbf{x}_{i,k}) = -A_i \mathbf{w}_{i,k} + \mathbf{d}_i$ and the optimal condition $\hat{\mathbf{1}}^T \nabla F(X^\star) = 0$ that $(X^\star - \bar{X}_k)^T (\mathbf{A} W_k - \mathbf{A} W^\star) = (\bar{\mathbf{x}}_k - \mathbf{x}^\star)^T \hat{\mathbf{1}}^T \nabla F(X_k)$ holds. Thus, we have

$$\begin{aligned}
&(X^\star - \bar{X}_k)^T (\mathbf{A} W_k - \mathbf{A} W^\star) \\
=&\frac{n}{\beta_{k-1}} (\bar{\mathbf{x}}_k - \mathbf{x}^\star)^T (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{k+1}) \\
&+ \frac{1}{\beta_{k-1}} (\bar{\mathbf{x}}_k - \mathbf{x}^\star)^T \left(\hat{\mathbf{1}}^T \eta_k \mathbf{n}_{X_k} - \hat{\mathbf{1}}^T \gamma_{k-1} \mathbf{n}_{L_{k-1}} \right),
\end{aligned} \tag{14}$$

where we use the update rule (11) and relation

$$\hat{\mathbf{1}}^T (L_k - L_{k-1}) = \hat{\mathbf{1}}^T (\beta_{k-1} \nabla F(X_k) + \gamma_{k-1} \mathbf{n}_{L_{k-1}}). \tag{15}$$

Taking expectation on both sides of (14) and using $2(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{c}) = \|\mathbf{a}-\mathbf{b}\|^2 + \|\mathbf{a}-\mathbf{c}\|^2 - \|\mathbf{b}-\mathbf{c}\|^2$ yields

$$\mathbb{E}[(X^\star - \bar{X}_k)^T(\mathbf{A}W_k - \mathbf{A}W^\star)] = \frac{1}{2\beta_{k-1}}\big(\mathbb{E}[\|\bar{X}_k - X^\star\|^2] + \mathbb{E}[\|\bar{X}_k - \bar{X}_{k+1}\|^2]\big) - \frac{1}{2\beta_{k-1}}\mathbb{E}[\|\bar{X}_{k+1} - X^\star\|^2]. \quad (16)$$

Using Cauchy-Schwarz inequality, we have $(\bar{X}_k - X_k)^T(\mathbf{A}W_k - \mathbf{A}W^\star) \le \frac{\|\mathbf{A}\|^2}{2\mu}\|X_k - \bar{X}_k\|^2 + \frac{\mu}{2}\|W_k - W^\star\|^2$. Combining this relation with (13), (16) yields

$$\begin{aligned} m_{k+1,1} \le & m_{k,1} + L'\beta_{k-1}m_{k,3} + \mathbb{E}[\|L_k - L_{k-1}\|^2] \\ & + \eta_k^2\sigma_X^2 - \mu\beta_{k-1}\mathbb{E}[\|W_k - W^\star\|^2], \end{aligned} \quad (17)$$

where we use $\mathbb{E}[\|\bar{X}_k - \bar{X}_{k+1}\|^2] \le \mathbb{E}[\|L_k - L_{k-1}\|^2] + \eta_k^2\sigma_X^2$.

From relation (12), we have

$$\begin{aligned} L_k - L_{k-1} = & \gamma_{k-1}(\boldsymbol{\mathcal{W}} - I)(L_{k-1} - \bar{L}_{k-1}) \\ & + \beta_{k-1}\nabla F(X_k) + \gamma_{k-1}\mathbf{n}_{L_{k-1}}. \end{aligned} \quad (18)$$

It follows from the L'-smoothness of $F(X)$ and $\hat{\mathbf{1}}^T\nabla F(X^\star) = 0$ that $F(\bar{X}_k) \ge F(X^\star) + \frac{1}{2L'}\|\nabla F(\bar{X}_k) - \nabla F(X^\star)\|^2$. Thus, we have $\|\nabla F(X_k)\|^2 \le 3L'^2\|X_k - \bar{X}_k\|^2 + 6L'\big(F(\bar{X}_k) - F(X^\star)\big) + 3\|\nabla F(X^\star)\|^2$. Combining this relation with (18) yields

$$\begin{aligned} & \mathbb{E}[\|L_k - L_{k-1}\|^2] \\ \le & 2\|\boldsymbol{\mathcal{W}} - I\|^2\gamma_{k-1}^2 m_{k,4} + 2\beta_{k-1}^2\mathbb{E}[\|\nabla F(X_k)\|^2] + \gamma_{k-1}^2\sigma_L^2 \\ \le & 12L'\beta_{k-1}^2 m_{k,2} + 6L'^2\beta_{k-1}^2 m_{k,3} + 2\|\boldsymbol{\mathcal{W}} - I\|^2\gamma_{k-1}^2 m_{k,4} \\ & + 6\beta_{k-1}^2\|\nabla F(X^\star)\|^2 + \gamma_{k-1}^2\sigma_L^2. \end{aligned} \quad (19)$$

Substituting (19) into (17) completes the proof. ■

Appendix II
Proof of Lemma 3

Using the L'-smoothness of $F(X)$, we have $F(\bar{X}_{k+1}) \le F(\bar{X}_k) + \nabla F(\bar{X}_k)^T\hat{\mathbf{1}}(\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k) + \frac{nL'}{2}\|\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k\|^2$, where $L' = \frac{\|\mathbf{A}\|^2}{\mu}$. It follows from relation (11) that $\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k = \frac{\hat{\mathbf{1}}^T}{n}\left((\boldsymbol{\mathcal{W}}_{\eta_k} - I)X_k - (L_k - L_{k-1}) + \eta_k\mathbf{n}_{X_k}\right) = \frac{\hat{\mathbf{1}}^T}{n}\left(-(L_k - L_{k-1}) + \eta_k\mathbf{n}_{X_k}\right)$. Combining these two relations and taking expectation on both sides yields

$$\begin{aligned} m_{k+1,2} \le & m_{k,2} + \frac{L'}{2}\mathbb{E}[\|L_k - L_{k-1}\|^2] + \frac{L'}{2}\eta_k^2\sigma_X^2 \\ & + \mathbb{E}[-\frac{1}{n}(\hat{\mathbf{1}}^T\nabla F(\bar{X}_k))^T\hat{\mathbf{1}}(L_k - L_{k-1})] \end{aligned} \quad (20)$$

Substituting (15) into the last term in (20) yields

$$\begin{aligned} & \mathbb{E}[-\frac{1}{n}(\hat{\mathbf{1}}^T\nabla F(\bar{X}_k))^T\hat{\mathbf{1}}(L_k - L_{k-1})] \\ \le & \frac{L'^2}{2}\beta_{k-1}m_{k,3} - \frac{1}{2n}\beta_{k-1}\mathbb{E}[\|\hat{\mathbf{1}}^T\nabla F(\bar{X}_k)\|^2]. \end{aligned} \quad (21)$$

where we use the Cauchy-Schwarz inequality.

Substituting (19), (21) into (20), we have

$$\begin{aligned} m_{k+1,2} \le & \frac{L'}{2}\eta_k^2\sigma_X^2 + \frac{L'}{2}\gamma_{k-1}^2\sigma_L^2 + (1 + 6L'^2\beta_{k-1}^2)m_{k,2} \\ & + (\frac{L'^2}{2}\beta_{k-1} + 3L'^3\beta_{k-1}^2)m_{k,3} + L'\|\boldsymbol{\mathcal{W}} - I\|^2\gamma_{k-1}^2 m_{k,4} \\ & + 3L'\beta_{k-1}^2\|\nabla F(X^\star)\|^2 - \frac{\beta_{k-1}}{2n}\mathbb{E}[\|\hat{\mathbf{1}}^T\nabla F(\bar{X}_k)\|^2]. \end{aligned} \quad (22)$$

From relation (11), we have $X_{k+1} - \bar{X}_{k+1} = (\boldsymbol{\mathcal{W}}_{\eta_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n})(X_k - \bar{X}_k) - (I - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n})(\beta_{k-1}\nabla F(X_k) - \gamma_{k-1}(\boldsymbol{\mathcal{W}} - I)(L_{k-1} - \bar{L}_{k-1}) + \eta_k\mathbf{n}_{X_k} + \gamma_{k-1}\mathbf{n}_{L_{k-1}})$, where we use $\hat{\mathbf{1}}^T\mathcal{W}_{\eta_k} = \hat{\mathbf{1}}^T$, $\mathcal{W}_{\eta_k}\hat{\mathbf{1}} = \hat{\mathbf{1}}$ and (18). Combining this relation with (19) and $\|I - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}\| = 1$ yields

$$\begin{aligned} m_{k+1,3} \le & (1 - \eta_k\lambda_{\mathcal{W}})m_{k,3} + \frac{2}{\eta_k\lambda_{\mathcal{W}}}\gamma_{k-1}^2\|\boldsymbol{\mathcal{W}} - I\|^2 m_{k,4} \\ & + \frac{2}{\eta_k\lambda_{\mathcal{W}}}\beta_{k-1}^2\mathbb{E}[\|\nabla F(X_k)\|^2] + \eta_k^2\sigma_X + \gamma_{k-1}^2\sigma_L^2 \\ \le & \frac{12L'}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k}m_{k,2} + (1 - \eta_k\lambda_{\mathcal{W}} + \frac{6L'^2}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k})m_{k,3} \\ & + \frac{2\|\boldsymbol{\mathcal{W}} - I\|^2}{\lambda_{\mathcal{W}}}\frac{\gamma_{k-1}^2}{\eta_k}m_{k,4} + \frac{6}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k}\|\nabla F(X^\star)\|^2 \\ & + \eta_k^2\sigma_X^2 + \gamma_{k-1}^2\sigma_L^2, \end{aligned} \quad (23)$$

where the first inequality uses $\|\mathbf{a} + \mathbf{b}\|^2 \le (1 + \epsilon)\|\mathbf{a}\|^2 + (1 + \epsilon^{-1})\|\mathbf{b}\|^2$ and $\|\boldsymbol{\mathcal{W}}_{\eta_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}\| = 1 - \eta_k\lambda_{\mathcal{W}} < 1$.

Similarly, from relation (12), we have $L_k - \bar{L}_k = (\boldsymbol{\mathcal{W}}_{\gamma_{k-1}} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n})(L_{k-1} - \bar{L}_{k-1}) + \beta_{k-1}(I - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n})\nabla F(X_k) + (I - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n})\gamma_{k-1}\mathbf{n}_{L_{k-1}}$, where we use $\hat{\mathbf{1}}^T\mathcal{W}_{\gamma_k} = \hat{\mathbf{1}}^T$, $\mathcal{W}_{\gamma_k}\hat{\mathbf{1}} = \hat{\mathbf{1}}$. Combining this relation with (19) yields

$$\begin{aligned} m_{k+1,4} \le & (1 - \gamma_{k-1}\lambda_{\mathcal{W}})m_{k,4} \\ & + \frac{1}{\gamma_{k-1}\lambda_{\mathcal{W}}}\beta_{k-1}^2\mathbb{E}[\|\nabla F(X_k)\|^2] + \gamma_{k-1}^2\sigma_L^2 \\ \le & \frac{6L'}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}m_{k,2} + \frac{3L'^2}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}m_{k,3} + \gamma_{k-1}^2\sigma_L^2 \\ & + (1 - \gamma_{k-1}\lambda_{\mathcal{W}})m_{k,4} + \frac{3}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}\|\nabla F(X^\star)\|^2. \end{aligned} \quad (24)$$

where we use $\|\boldsymbol{\mathcal{W}}_{\gamma_k} - \frac{\hat{\mathbf{1}}\hat{\mathbf{1}}^T}{n}\| = 1 - \gamma_k\lambda_{\mathcal{W}} < 1$.

Combining relations (22), (23), (24) with Lemma 1 yields relation (7), where all matrices has nonnegative elements as

$$B_k = \begin{bmatrix} 1 & 0 & b_k^{1,3} & 0 \\ 0 & 1 & b_k^{2,3} & 0 \\ 0 & 0 & b_k^{3,3} & b_k^{3,4} \\ 0 & 0 & 0 & b_k^{4,4} \end{bmatrix},\ C_k = \begin{bmatrix} 0 & c_k^{1,2} & c_k^{1,3} & c_k^{1,4} \\ 0 & c_k^{2,2} & c_k^{2,3} & c_k^{2,4} \\ 0 & c_k^{3,2} & c_k^{3,3} & 0 \\ 0 & c_k^{4,2} & c_k^{4,3} & 0 \end{bmatrix},$$

$$D_k = \begin{bmatrix} d_k^1 & d_k^2 & d_k^3 & d_k^4 \end{bmatrix}^T,\ H_k = \begin{bmatrix} \mu\beta_{k-1} & 0 & 0 & 0 \\ 0 & \frac{\beta_{k-1}}{2n} & 0 & 0 \end{bmatrix}^T,$$

where $b_k^{1,3} = \frac{\|\mathbf{A}\|^2}{\mu}\beta_{k-1}$, $b_k^{2,3} = \frac{\|\mathbf{A}\|^4}{2\mu^2}\beta_{k-1}$, $b_k^{3,3} = 1 - \eta_k\lambda_{\mathcal{W}}$, $b_k^{3,4} = \frac{2\|\boldsymbol{\mathcal{W}} - I\|^2}{\lambda_{\mathcal{W}}}\frac{\gamma_{k-1}^2}{\eta_k}$, and $b_k^{4,4} = 1 - \gamma_{k-1}\lambda_{\mathcal{W}}$; $c_k^{1,2} = \frac{12\|\mathbf{A}\|^2}{\mu}\beta_{k-1}^2$, $c_k^{1,3} = \frac{6\|\mathbf{A}\|^4}{\mu^2}\beta_{k-1}^2$, $c_k^{1,4} = 2\|\boldsymbol{\mathcal{W}} - I\|^2\gamma_{k-1}^2$, $c_k^{2,2} = \frac{6\|\mathbf{A}\|^4}{\mu^2}\beta_{k-1}^2$, $c_k^{2,3} = \frac{3\|\mathbf{A}\|^6}{\mu^3}\beta_{k-1}^2$, $c_k^{2,4} = \frac{\|\mathbf{A}\|^2\|\boldsymbol{\mathcal{W}} - I\|^2}{\mu}\gamma_{k-1}^2$, $c_k^{3,2} = \frac{12\|\mathbf{A}\|^2}{\mu\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k}$, $c_k^{3,3} = \frac{6\|\mathbf{A}\|^4}{\mu^2\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k}$, $c_k^{4,2} = \frac{6\|\mathbf{A}\|^2}{\mu\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}$, and $c_k^{4,3} = \frac{3\|\mathbf{A}\|^4}{\mu^2\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}$; $d_k^1 = 6\beta_{k-1}^2\|\nabla F(X^\star)\|^2 + \eta_k^2\sigma_X^2 + \gamma_{k-1}^2\sigma_L^2$, $d_k^2 = \frac{\|\mathbf{A}\|^2}{2\mu}\big(6\beta_{k-1}^2\|\nabla F(X^\star)\|^2 + \eta_k^2\sigma_X^2 + \gamma_{k-1}^2\sigma_L^2\big)$, $d_k^3 = \frac{6}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\eta_k}\|\nabla F(X^\star)\|^2 + \eta_k^2\sigma_X^2 + \gamma_{k-1}^2\sigma_L^2$, and $d_k^4 = \frac{3}{\lambda_{\mathcal{W}}}\frac{\beta_{k-1}^2}{\gamma_{k-1}}\|\nabla F(X^\star)\|^2 + \gamma_{k-1}^2\sigma_L^2$. The proof is completed. ■