SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2018, Vol. 61 092202:1–092202:14
https://doi.org/10.1007/s11432-016-9173-8

Distributed regression estimation with incomplete data in multi-agent networks

Yinghui WANG[1,2*] , Peng LIN[1,2] & Yiguang HONG[2]

[1]*University of Chinese Academy of Sciences, Beijing* 100049*, China;*
[2]*Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China*

Received 23 November 2016/Accepted 21 June 2017/Published online 4 January 2018

Abstract In this paper, distributed regression estimation problem with incomplete data in a time-varying multi-agent network is investigated. Regression estimation is carried out based on local agent information with incomplete in the non-ignorable mechanism. By virtue of gradient-based design and adaptive filter, a distributed algorithm is proposed to deal with a regression estimation problem with incomplete data. With the help of convex analysis and stochastic approximation techniques, the exact convergence is obtained for the proposed algorithm with incomplete data and a jointly-connected multi-agent topology. Moreover, online regret analysis is also given for real-time learning. Then, simulations for the proposed algorithm are also given to demonstrate how it can solve the estimation problem in a distributed way, even when the network configuration is time-varying.

Keywords multi-agent systems, time-varying network, estimation with incomplete data, online learning, stochastic approximation

Citation Wang Y H, Lin P, Hong Y G. Distributed regression estimation with incomplete data in multi-agent networks. Sci China Inf Sci, 2018, 61(9): 092202, https://doi.org/10.1007/s11432-016-9173-8

1 Introduction

Recent years have witnessed a rapid development of multi-agent control and optimization [1–3], in order to break through the bottleneck confronted by conventional centralized design. Note that conventional parallel estimation or optimization in a network also requires a central unit to collect all the network data and then assign jobs to the network. Therefore, the distributed design, based on the local computation with inter-agent communication and the information source of each agent, becomes more and more important in many areas, including smart grids, communication networks, and economical systems [1,4–6].

In practice, uncertainties and time-varying structures are often concerned in distributed design. Incomplete date is quite common to deal with in many practical or observational situations such as targeted surveys and clinical trials [7, 8]. In fact, data are missing due to various reasons, including drop-out of the survey, time-delay of collecting data, and refusal to response to some items of the experimental investigations by participants. Note that the failure to address issues of incomplete data can give rise to the bias of the estimates of unknown parameters. As pointed out in [9], incomplete data are not ignorable. Because the probability of missing data lies on the values of unobserved variables, the non-ignorable missing case (also called the missing not at random case in [10]) is very general in many situations and frequently encountered in various research areas, including industrial data missing for occasional device

* Corresponding author (email: wangyinghuisdu@163.com)

misfunction and printing errors of a testing booklet for a subset of study participants [11]. In fact, the discussion on different types of missing data analysis can be easily found in the literature (see [9]).

Switching topologies between agents bring about challenges in the convergence study of multi-agent systems, where the time-varying interactions often result from the energy-saving policy or link failure. Actually, significant efforts have been made for distributed optimization problems with switching multi-agent topologies such as jointly-connected networks [1, 5, 12]. Among these research results, variable connectivity of networks plays a vital part in the exploration of multi-agent coordination, which makes distributed algorithms/methods suitable or effective in dealing with large data or complicated network structures. In fact, distributed optimization results may be extended to these situations, considering that regression estimation can be converted to an optimization problem.

Regression can be used in data-based estimation and optimization with many practical backgrounds, and therefore, various regression models have been proposed [13]. Among the models, the linear regression model is still the most popular one owing to its simplicity and efficiency, and its distributed design becomes more and more popular [10]. On the other hand, online learning also becomes a hot topic to handle the complicated and time-varying situations [14]. Online regression learning measures how well the regression model can predict the unknown parameters associated with unobserved regressors. The regret or estimation error [15] in regression estimation is defined as the difference between the total cost achieved by the regression model and that of the best fixed regression model in hindsight. Since the regret function for the estimation error during a period of time is often convex, online learning is closely related to online (convex) optimization. In the strong convexity case, the lower bounds for the regret functions was achieved with $O(\frac{\log(T)}{T})$ in [16], and then some analyses were extended to distributed online computation (e.g., classification or learning with privacy preserving in [14, 17]).

The motivation of our research in the paper is for efficient regression estimation algorithms/techniques with large data and time-varying networks. Here we consider distributed computation for the estimation in a time-varying jointly-connected network with non-ignorable missing data. In the network, each agent is able to obtain a stream of incomplete regression data. The agents need to communicate with others to calculate the unknown parameters in the regression estimation problem.

The technical contribution of the paper is summarized as follows. (i) We propose a distributed design for a time-varying network, where all the computation in the multi-agent network has to be carried out by local information. Different from [10] for the regressor analysis with the missing completely at random mechanism, we consider the non-ignorable missing data mechanism. Both theoretical and numerical analyses are provided to demonstrate the effectiveness of the given distributed design ground on the inter-agent communication and local agent information. (ii) By combining the idea of distributed (sub)gradient method given in [1, 12] along with the adaptive filter idea in [18], we design a distributed estimation algorithm which can adapt to the streaming incomplete data. Different from the confusion strategy given in [10, 19, 20], our algorithm based on diminishing step-size can get an exact solution. (iii) We also provide distributed online learning scheme for real-time computation in the non-ignorable missing data case and give the regret analysis to measure the effectiveness of the given algorithm. As far as we know, this is the first effort for the regret analysis about distributed online learning with missing data in the non-ignorable missing case, though there are few distributed online algorithms for some other learning problems (referring to [14, 17]).

The outline of the paper is given as follows. Some necessary preliminaries about convex analysis and graph theory are introduced in Section 2, while the distributed estimation problem with incomplete data in the non-ignorable missing case is formulated and a distributed adaptive gradient algorithm (DAGA) is presented in Section 3. Next, the online regret analysis of the distributed adaptive gradient algorithm (DAGA) is given in Section 4, and then simulation examples are shown in Section 5. Finally, the concluding remarks are given in Section 6.

2 Preliminaries

In order to formulate the distributed regression estimation problem, in this section, we first offer some necessary preliminary knowledge related to convex analysis [21] and graph theory [22].

2.1 Convex analysis

The following concepts about convex function and projection operations are detailed in [21]. A differentiable function $f : \mathbf{R}^m \to \mathbf{R}$ is called μ-strongly convex on $\mathbf{R}^m$ if, for any $x_1, x_2 \in \mathbf{R}^m$, the gradient $\nabla f(\cdot)$ of $f(\cdot)$ satisfies $f(x_2) \geqslant f(x_1) + \nabla^{\mathrm{T}} f(x_1)(x_2 - x_1) + \frac{1}{2}\mu\|x_2 - x_1\|^2$, where μ is a constant. It is called to have a locally Lipschitz continuous gradient, if, for any given compact set Λ (*i. e.*, Λ is a bounded closed set in $\mathbf{R}^m$), there is a constant k_Λ such that

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leqslant k_\Lambda \|x - y\|, \quad \forall x_1, x_2 \in \Lambda. \tag{1}$$

Denote $P_X(x) = \arg\min_{y \in X} \|x - y\|$ as the projection from x onto X, where X is a closed convex set. Obviously, $\langle x_1 - P_X(x_1), x_2 - P_X(x_1)\rangle \leqslant 0$, for all $x_1 \in \mathbf{R}^m$ and for all $x_2 \in X$. Then in light of the projection property, we get

$$\|x_1 - P_X(x_1)\|^2 + \|x_2 - P_X(x_1)\|^2 \leqslant \|x_1 - x_2\|^2, \quad \forall x_1 \in \mathbf{R}^m,\ \forall x_2 \in X, \tag{2}$$

and $\|P_X(x_1) - P_X(x_2)\| \leqslant \|x_1 - x_2\|$, $\forall x_1, x_2 \in \mathbf{R}^m$.

Denote $\{x_n\}$ as a sequence of vectors, and x is a fixed vector. $\mathbb{E}\|x\| \leqslant \infty$. If $\mathbb{E}\|x_n - x\| \to 0$ as $n \to \infty$, we say the sequence $\{x_n\}$ converge to vector x in L_1 space [23].

2.2 Graph theory

Denote a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ as the information sharing or exchanging network among N agents, where $\mathcal{N} = 1, 2, \ldots, N$ is the agent set of $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ is the edge set standing for the information communication among the N agents. $(j, i) \in \mathcal{E}$ means a directed path form j to i, which implies that agent i can receive information from its neighbor j. $\mathcal{N}_i = \{j | (j, i) \in \mathcal{E}\}$ indicates agent i's one-hop neighbor set. A path of $\mathcal{G}$ is a sequence of edges in $\mathcal{G}$ which connect a sequence of distinct agents in $\mathcal{N}$. If there is a path from agent j to i, j is called to be connected to i. If there is a directed path from i to j for any $i, j \in \mathcal{G}$, $\mathcal{G}$ is called to be strongly connected.

Consider a directed network $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k))$, which represents the (time-varying) information exchange of a multi-agent network at time k. Let $W(k) = [w_{ij}(k)]_{ij}$ (where $w_{ij}(k) \neq 0 \Leftrightarrow (j(k), i(k)) \in \mathcal{E}(k)$) denote the adjacency matrix $W(k) = [w_{ij}(k)]_{ij}$ of $\mathcal{G}$. Denote $\mathcal{E}_\infty = \{(j, i) : (j, i) \in \mathcal{E}_k,\ i.\ o.\}$, which means that $\mathcal{E}_\infty$ consists of edge pairs (j, i) such that agent j can send information to i infinitely often. The topology of the network $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k))$ is uniformly strongly jointly connected if, for all k, the union graph $(\mathcal{N}, \cup_{l=1,2,\ldots,\kappa} \mathcal{E}(k + l))$ is strongly connected and a given integer $\kappa > 0$, which the union graph is connected for every period κ.

The following assumptions are about the communication topology $\mathcal{G}(k)$ and its adjacency matrix $W(k)$ (see [1, 12]).

Assumption 1. There exists a constant c with $0 < c < 1$ such that, $\forall k \geqslant 0$ and $\forall i \in \{1, 2, \ldots, N\}$,
(1) $w_{ii}(k) \geqslant c$;
(2) $w_{ij}(k) \geqslant c$ if agent $j \in \mathcal{N}_i(k)$ and $w_{ij}(k) = 0$ otherwise;
(3) W is doubly stochastic.

Assumption 2. The network topology $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k))$ is uniformly strongly jointly connected, and moreover, there is an integer $\kappa \geqslant 1$ such that $\forall k \geqslant 0$ and $\forall (j, i) \in \mathcal{E}_\infty$,

$$(j, i) \in \mathcal{E}(k) \cup \mathcal{E}(k + 1) \cup \cdots \cup \mathcal{E}(k + \kappa - 1). \tag{3}$$

Assumption 2 provides a quite general connectivity condition for the multi-agent network, which was widely used in [1,12] . Moreover, from (3), it is easy to see that each agent i can receive information from all the neighbors in N_i at least once during each period of κ, though the network topology is time-varying and may be disconnected at each time.

3 Problem formulation and algorithm

In this section, we formulate our regression estimation problem and then propose a distributed algorithm.

3.1 Problem formulation

Roughly speaking, for a network of N individual agents, our problem is to estimate the parameter vector $(\xi^i)^* \in \mathbf{R}^M$ for agent i by using measurement incomplete data collected by these agents in the non-ignorable missing case.

Let us first check a simple case with complete data. Denote A_k as the complete data matrix and y_k as the corresponding regression vector that the network get at time k. A_k^i is the ith column of matrix A_k and y_k^i, the ith component of y_k, is a real number. Each agent i can obtain a stationary data sequence $\{A_k^i, y_k^i\}$ for $k = 1, 2, \ldots$, where $\{A_k^i\} \in \mathbf{R}^M$ are i.i.d. regression vector sequences and $\{y_k^i\}$ are observation sequences, and moreover, the sequences for different agents are independent of each other. Suppose that the data can be observed by agent i via a linear regression model as follows:

$$y_k^i = (A_k^i)^{\mathrm{T}}\xi^i + b_k^i, \quad i = 1, \ldots, N, \tag{4}$$

where $\{b_k^i\}$ is an i.i.d. zero-mean white noise process with variance $\sigma_{b,i}^2$. For agent i, A_k^i and y_k^i satisfy the ergodicity property [24]:

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}A_k^i(A_k^i)^{\mathrm{T}} = R^{A,i}, \quad \lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}y_k^iA_k^i = r^{yA,i}, \tag{5}$$

where $R^{A,i} = \mathbb{E}A_k^i(A_k^i)^{\mathrm{T}}$ stands for the covariance matrix of vector A_k^i and $r^{yA,i} = \mathbb{E}y_k^iA_k^i$. The ergodicity of A_k^i and y_k^i implies (5) with probability 1. From linear regression model (4), based on agent i's complete data, we give the minimum mean-square-error (MSE) estimate of $(\xi^i)^*$ for agent i and the corresponding MSE cost $f^i(\xi)$: $(\xi^i)^* = (R^{A,i})^{-1}r^{yA,i}$ and $f_i(x) = \mathbb{E}\|y_k^i - (A_k^i)^{\mathrm{T}}\xi\|^2$. The next assumption about the regression vectors $\{A_k^i\}$ and the noise processes $\{b_k^i\}$ is also widely-used (see [10]).

Assumption 3. The regression vectors $\{A_k^i\}$ and the noise processes $\{b_k^i\}$ are mutually independent of each other, and moreover, the covariance matrix $R^{A,i} = \mathbb{E}A_k^i(A_k^i)^{\mathrm{T}}$ is diagonal.

Next, let us go back to study the non-ignorable missing data mechanism. To deal with the mechanism, here we use single imputation methods. Namely, if we have some prior knowledge to the incomplete data, then we can replace the missing data by some deterministic or random values rather than delete them. Through single imputation methods, we can still get a complete data set. Next, we give a distributed design for the non-ignorable missing case. We express the incomplete data of regression vector A_k^i in the following form: $\bar{A}_k^i = \mathrm{col}\{m_{i1}(A_k^{i1}), m_{i2}(A_k^{i2}), \ldots, m_{iM}(A_k^{iM})\}$, where $m_{ij}(A_k^{ij}) = A_k^{ij}\mathbb{I}(|A_k^{ij}| \geqslant \theta^{ij})$ and $\theta^{ij} > 0$ is a predefined threshold (referring to [25]).

Remark 1. Assumption 3 also implies that the incomplete regression vectors $\{\bar{A}_k^i\}$ and the noise processes $\{b_k^i\}$ are mutually independent of each other. Moreover,the covariance matrix $R^{\bar{A},i} = \mathbb{E}\bar{A}_k^i(\bar{A}_k^i)^{\mathrm{T}}$ of the regression vector $\bar{A}_k^i$ is diagonal.

Agent i's minimum mean-square-error (MSE) estimate $(\xi^i)^o$ based on the data $\{\bar{A}_k^i, y_k^i\}$ is

$$(\xi^i)^o = (R^{\bar{A},i})^{-1}r^{y\bar{A},i}. \tag{6}$$

The corresponding MSE cost is $f_{\mathrm{missing}}^i(\xi) = \mathbb{E}\|y_k^i - (\bar{A}_k^i)^{\mathrm{T}}\xi\|^2$. With the non-ignorable missing data mechanism, each element $\bar{A}_k^{ij}$ of $\bar{A}_k^i$ is independent of each other, and

$$R^{\bar{A},i} = \mathrm{diag}\{\mathbb{E}\bar{A}_k^{i1}(\bar{A}_k^{i1})^{\mathrm{T}}, \mathbb{E}\bar{A}_k^{i2}(\bar{A}_k^{i2})^{\mathrm{T}}, \ldots, \mathbb{E}\bar{A}_k^{iM}(\bar{A}_M^{i1})^{\mathrm{T}})\}.$$

Therefore, the covariance matrix $R^{\bar{A},i}$ of the data-missing regressor $\bar{A}_k^i$ becomes

$$R^{\bar{A},i} = \int_\Omega \bar{A}_k^{ij}(\bar{A}_k^{ij})^{\mathrm{T}}\mathrm{d}P = \int_{|A_k^{ij}|\geqslant\theta^{ij}} A_k^{ij}(A_k^{ij})^{\mathrm{T}}\mathrm{d}P = \tau_{ij}\int_\Omega A_k^{ij}(A_k^{ij})^{\mathrm{T}}\mathrm{d}P = \tau_{ij}R^{A,i},$$

where $0 < \tau_{ij} \leqslant 1$. Define $\tau_i = \mathrm{diag}\{\tau_{i1}, \tau_{i2}, \ldots, \tau_{iM}\}$, and then $R^{\bar{A},i} = \tau_iR^{A,i}$. By analogous analysis, $r^{y\bar{A},i} = \tau_ir^{yA,i}$. Therefore, the estimate of the unknown vector $(\xi^i)^o$ in the non-ignorable missing case is the same with the estimate of the complete data, i.e., $(\xi^i)^o = (R^{\bar{A},i})^{-1}r^{y\bar{A},i} = (\tau_iR^{A,i})^{-1}\tau_ir^{yA,i} = (\xi^i)^*$.

Remark 2. If we choose $m_{ij}(A_k^{ij}) = A_k^{ij}\mathbb{I}(|A_k^{ij}| \leqslant \theta^{ij})$ for data imputation, then we can still prove $(\xi^i)^o = (\xi^i)^*$ in a similar way.

Thus, the distributed estimation problem considered in this paper is to make agent i be able to find the common $(\xi^i)^*$ that minimizes the mean-square-error function $f^i_{\text{missing}}(x)$ based on a time-varying inter-agent topology $\mathcal{G}(k) = (\mathcal{N}, \mathcal{E}(k))$. To be strict, if we can find ξ^* to achieve

$$\min \ \ g(\xi) = \sum_{i=1}^{N} g^i(\xi) = \sum_{i=1}^{N} f^i_{\text{missing}}(\xi), \ \text{s. t.}\ \xi \in X. \tag{7}$$

where $X \subset R^M$ is compact and convex, and $x \in X$ guarantees that $\xi^i,\ \forall i \in \mathcal{N}$ is bounded even when the network topology is time-varying.

Remark 3. Since $X \subset R^M$ is compact and convex for all $\xi \in X$, $\|\xi\| \leqslant C_x$. Moreover, the cost function $g^i(\xi)$ is μ-strongly convex because we can prove that $g^i(\xi)$ satisfies (1) by a simple calculation.

3.2 Distributed adaptive gradient-based algorithm

Here we propose a distributed design to recover ξ^* by using an alternative function $f^i_{\text{missing}}(\xi)$ in the non-ignorable missing data mechanism, when each agent can only get information from its one-hop neighbors and itself. To be specific, we propose the following distributed gradient-based design:

$$\begin{cases} \hat{\xi}^i_{k+1} = \sum\limits_{j=1}^{n} w_{ij}(k)\xi^i_k - \iota_k d^i_k, \\ \xi^i_{k+1} = P_X(\hat{\xi}^i_{k+1}), \end{cases} \tag{8}$$

where ι_k is a diminishing step-size satisfying the stochastic approximation conditions [26]:

$$\iota_k > 0, \ \ \lim_{k\to\infty} \iota_k = 0, \tag{9}$$

$$\iota_k > 0, \ \ \sum_{i=1}^{\infty} \iota_k = \infty, \ \ \sum_{i=1}^{\infty} \iota_k^2 < \infty. \tag{10}$$

Define $\triangledown g^i_k$ as the gradient of function $g^i(\xi)$ at ξ^i_k and $d^i_k = \triangledown g^i_k + \epsilon^i_k$ as the disturbed gradient of function $\triangledown g^i(\xi)$, where ϵ^i_k represents the observation noise. Agent i can generate a $\mathbf{R}^M$-valued stochastic process due to the above distributed gradient-based design.

We also need to estimate the parameters in the distributed design. The gradient information $\triangledown g_i(\xi^i_k)$ at point ξ^i_k requires the knowledge of $R^{\bar{A},i}$ and $r^{y\bar{A},i}$. To deal with the challenge, we estimate $R^{\bar{A},i}$ and $r^{y\bar{A},i}$ using adaptive filters [18, 27] for data exploration:

$$R^{\bar{A},i}_k = (1-\rho_k)R^{\bar{A},i}_{k-1} + \rho_k \bar{A}^i_k(\bar{A}^i_k)^{\mathrm{T}}, \tag{11}$$

$$r^{y\bar{A},i}_k = (1-\rho_k)r^{y\bar{A},i}_{k-1} + \rho_k y^i_k \bar{A}^i_k, \tag{12}$$

where $\rho_k \to 0$ as $k \to \infty$.

The structure of adaptive filter used here is alterable and adjustable so that the behavior or performance of $R^{\bar{A},i}_k$ and $r^{y\bar{A},i}_k$ can improve through contact with previous information of $\bar{A}^i_k$ and y^i_k. Several strong points about the usage of adaptive filters were clearly pointed out in [18], including the automatic adaptation for changing environments and changing system requirements, self-design without requiring the elaborate synthesis procedures, and extrapolation of a model for new situations.

Under Assumption 3 along with the ergodic property of $\bar{A}^i_k$, we conclude that $\bar{A}^i_k$ is also ergodic, and by induction, we have $\lim_{k\to\infty} R^{\bar{A},i}_k = \lim_{k\to\infty} R^{\bar{A},i}_{k-1} + \rho_k(\bar{A}^i_k(\bar{A}^i_k)^{\mathrm{T}} - R^{\bar{A},i}_{k-1}) = R^{\bar{A},i}$. Analogously, by the stationary property of y^i_k and $\bar{A}^i_k$, we obtain $\lim_{k\to\infty} r^{y\bar{A},i}_k = \lim_{k\to\infty} r^{y\bar{A},i}_{k-1} + \rho_k(y^i_k\bar{A}^i_k - r^{y\bar{A},i}_{k-1}) = r^{y\bar{A},i}$.

Next, we propose our Distributed adaptive gradient-based algorithm (DAGA) in the non-ignorable missing data mechanism in Algorithm 1.

Algorithm 1 Distributed adaptive gradient-based algorithm (DAGA)

1: $R_k^{\bar{A},i} = (1-\rho_k)R_{k-1}^{\bar{A},i} + \rho_k \bar{A}_k^i (\bar{A}_k^i)^{\mathrm{T}}$;
2: $r_k^{y\bar{A},i} = (1-\rho_k)r_{k-1}^{y\bar{A},i} + \rho_k y_k^i \bar{A}_k^i$;
3: $d_k^i = R_k^{\bar{A},i}\xi_k^i - r_k^{y\bar{A},i}$;
4: $\hat{\xi}_{k+1}^i = \sum_{j=1}^N w_{ij}(k)\xi_k^i - \iota_k d_k^i$;
5: $\xi_{k+1}^i = P_X(\hat{\xi}_{k+1}^i)$.

Note that this gradient-based algorithm is distributed because there is no central unit in the network. Instead, each agent has to combine information gathered from its neighbors and with its local sources in order to achieve global optimization. Furthermore, different from the algorithms given in [1, 12], the algorithm is also "adaptive" because we estimate the gradient by means of adaptive filters.

Remark 4. In fact, we can rewrite the DAGA (8) as follows:

$$\xi_{k+1}^i = \sum_{j=1}^N w_{ij}(k)\xi_k^i - \iota_k(\nabla g_k^i + \epsilon_k^i). \tag{13}$$

In (13), $\nabla g_k^i = R^{\bar{A},i}\xi_k^i - r^{y\bar{A},i}$ is the gradient of function $g^i(\xi)$ and $\epsilon_k^i = d_k^i - \nabla g_k^i = R_k^{\bar{A},i}\xi_k^i - R^{\bar{A},i}\xi_k^i + r^{y\bar{A},i} - y_k^i\bar{A}_k^i$ is the observation noise.

Remark 5. Under Remark 3, the cost function $g^i(\xi)$ is strongly convex. Still, $\|\nabla g^i(\xi) - \nabla g^i(y)\| = 2\bar{A}_k^i(\bar{A}_k^i)^{\mathrm{T}}(\xi - y)$, satisfies the locally Lipschitz continuous condition. Therefore, in the convex set X, the gradient $\nabla g^i(\xi)$ of $g^i(\xi)$ is uniformly bounded, i.e., $\|\nabla g^i(\xi)\| \leqslant C_g,\ \forall x \in X$. Also, it follows from [21] that $\forall \xi, y \in X,\ \ g^i(y) - g(\xi) - \langle y - \xi,\ \ \nabla g^i(\xi)\rangle \geqslant 0$.

4 Main results of DAGA

In this section, convergence and related online performance of Algorithm 1 are strictly analyzed. The proofs of some lemmas and theorems can be found in Appendix.

4.1 Convergence analysis of DAGA

In this subsection, we give the convergence analysis step by step. First, we prove that all the agents achieve consensus in L_1 space and almost surely. Then we demonstrate that each agent reaches the same optimal solution of (7) almost surely.

Denote the transition matrix of $W(k)$ as $\Psi(k,s) = W(k)W(k-1)\cdots W(s), k \geqslant s$, where $[\Psi(k,s)]_{ij}$ stands for the ijth element of $\Psi(k,s)$. The following lemma is about $\Psi(k,s)$, given in Proposition 1 of [1].

Lemma 1. Under Assumptions 1-2, $\|[\Psi(k,s)]_{ij} - \frac{1}{N}\| \leqslant \lambda\beta^{k-s},\ \forall k > s$, where $\lambda = 2(1+\eta^{-K_0})/(1-\eta^{-K_0})$, with $K_0 = (m-1)\kappa$, $\beta = (1-\eta^{-K_0})^{1/K_0}$.

Define $F_k = \sigma\{\bar{A}_k^i, y_k^i, i \in \mathcal{N}, 1 \leqslant l \leqslant k\}$ for all $k \geqslant 1$, where F_k is the σ-algebra created by the whole history of DAGA up to moment k (referring to [28] and [19]).

The following result can be found in [21], which is useful for the convergence analysis of Algorithm 1.

Lemma 2. Let $(\Omega, F, \mathbb{P})$ denote a probability space. F_n is a filtration, which is an increasing sequence of F. $\{d_k\}$,$\{v_k\}$ and $\{w_k\}$ are F_k-measurable scalar random variables in $(\Omega, F, \mathbb{P})$. $\{v_k\}$ and $\{w_k\}$ are both nonnegative with $\sum_{k=1}^\infty w_k < \infty$ and $\{d_k\}$ is bounded below uniformly. If

$$E[d_{k+1}|F_k] \leqslant (1+\eta_k)d_k - v_k + w_k,\ \ \forall k \geqslant 1 \tag{14}$$

holds with probability 1, where $\eta_k \geqslant 0$ are constants with $\sum_{k=1}^\infty \eta_k < \infty$, then $\{d_k\}$ converges almost surely with $\sum_{k=1}^\infty v_k < \infty$.

Lemma 3 shows the boundedness of d_k^i and the observation noise ϵ_k^i between ∇g_k^i and d_k^i, whose proof is in Appendix A.

Lemma 3. With Assumptions 1-3 hold, for all $k \geqslant 0$, $\mathbb{E}\|\epsilon_k^i\|$ and $\mathbb{E}\|d_k^i\|$ are both bounded.

Let us consider the difference between ξ_k^i and $\bar{\xi}_k = \frac{1}{N}\sum_{i=1}^N \xi_k^i$ in the following result, whose proof is in Appendix B.

Lemma 4. With Assumptions 1–3 and (10),

$$\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| \leqslant N\lambda\beta^k \max_{1\leqslant j\leqslant N}\|\xi_0^j\| + \iota_k\|d_k^i\| + \frac{\iota_k}{N}\sum_{i=1}^N\|d_k^i\| + \lambda\sum_{s=1}^k\beta^{k-s}\sum_{i=1}^N\iota_{s-1}\|d_{s-1}^i\| \tag{15}$$

for all agent i, and $k \geqslant 0$.

Next, we show that the agents in the network can reach consensus in L_1 space by Algorithm 1.

Theorem 1. With Assumptions 1-3 and (10), the consensus in L_1 space can be achieved by Algorithm 1, that is, for $i, j = 1, 2, \ldots, N$, $\lim_{k\to\infty}\mathbb{E}[\|\xi_k^i - \xi_k^j\|] = 0$.

Proof. Taking the expectation of (15), the following inequality holds:

$$\mathbb{E}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| \leqslant N\lambda\beta^k \max_{1\leqslant j\leqslant N}\|\xi_0^j\| + \iota_k\mathbb{E}\|d_k^i\| + \frac{\iota_k}{N}\sum_{i=1}^N\mathbb{E}\|d_k^i\| + \lambda\sum_{s=1}^k\beta^{k-s}\sum_{i=1}^N\iota_{s-1}\mathbb{E}\|d_{s-1}^i\|. \tag{16}$$

From Lemma 3, $\mathbb{E}\|d_k^i\| \leqslant M_d$. Therefore, $\mathbb{E}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| \leqslant N\lambda\beta^k C_x + 2\iota_k M_d + \lambda N M_d\sum_{s=1}^k\iota_{s-1}\beta^{k-s}$. Due to the diminishing step-size ι_k, $\lim_{k\to\infty}\sum_{s=1}^k\iota_{s-1}\lambda\beta^{k-s} = 0$. Therefore, $\lim_{k\to\infty}\mathbb{E}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| = 0$, $\forall i \in \mathcal{N}$.

Moreover, we demonstrate that all the agents $i,\ i \in \mathcal{N}$ achieve consensus almost surely, whose proof can be found in Appendix C.

Theorem 2. With Assumptions 1–3 and (10), the consensus almost surely can be achieved by Algorithm 1, that is, for $i, j = 1, 2, \ldots, N$, the following equality $\lim_{k\to\infty}\|\xi_k^i - \xi_k^j\| = 0$ almost surely holds.

The following theorem is on the convergence to the optimal solution almost surely of DAGA, whose proof is in Appendix D.

Theorem 3. With Assumptions 1–3 and (10), all the sequences $\{\xi_k^i\},\ i \in \mathcal{N}$ converge to the optimal point ξ^* almost surely by DAGA.

4.2 Online analysis of DARA

Based on the above results, we go further to analyze online performance based on our algorithm. To be specific, we show our regret bound in the non-ignorable missing case, which is consistent with the conclusions in the online convex optimization.

In the online learning setup, agent i can not only use partial data $\bar{A}^i$ and y^i to update its state ξ^i, but also communicate its state ξ^i with its one-hop neighbors through communication topologies described in Subsection 2.2. In this subsection, we give a corresponding online learning scheme, which can be formulated as an online optimization problem as follows:

$$\min\ \ G(\xi) = \sum_{k=1}^{\mathrm{T}}\sum_{i=1}^N g_k^i(\xi) \tag{17}$$

for a given constant $T > 0$. Here, we introduce a regret function to measure the regret or estimation error, which has been widely used in online optimization (referring to [14]). In fact, the (average) regret function during the time interval $[1, T]$ can be denoted as

$$R(T) = \frac{1}{T}\mathbb{E}\left[\sum_{k=1}^{\mathrm{T}} g(\xi_k) - \sum_{k=1}^{\mathrm{T}} g(\xi^*)\right], \tag{18}$$

where ξ^* stands for the optimal solution of the online circumstance. As usual, $R(T) = \frac{1}{T}R(T)$ stands for the (average) estimation error as used in [15]. Moreover, we still denote $\bar{\xi}_k = \frac{1}{N}\sum_{i=1}^N\xi_k^i$ as the average parameter of vector $\xi_k^i, i = 1, 2, \ldots N$.

Lemma 5 is about the decomposition of the regret function, whose proof is in Appendix E.

Lemma 5. Let $\{\xi_k^i\}$ denote the sequences generated by DAGA. Define $\nabla\bar{g}_k^i$ as the gradient of $g^i(\xi)$ at $\bar{\xi}_k$. Then

$$\|\bar{\xi}_{k+1}-\xi\|^2 \leqslant \frac{9\iota_k^2}{N^2}\left(\sum_{i=1}^N\|d_k^i\|\right)^2 - \frac{2\iota_k}{N}(g(\xi)-g(\bar{\xi}_k)) + \frac{2\iota_k}{N}\sum_{i=1}^N(\|\nabla g_k^i\|+\|\nabla\bar{g}_k^i\|+\|\epsilon_k^i\|)\|\bar{\xi}_k-\xi_k^i\| + \frac{2\iota_k}{N}\sum_{i=1}^N\|\epsilon_k^i\|\|\xi_k^i-\xi\| + \frac{4\iota_k}{N}\sum_{i=1}^N\|d_k^i\|\|\bar{\xi}_k-\hat{\xi}_{k+1}^i\| + (1-\frac{\mu\iota_k}{2})\|\bar{\xi}_k-\xi\|^2. \tag{19}$$

Next we give the bound of the regret function (18) without using step-size condition (10).

Theorem 4. With Assumptions 1–3 and (9), $\sum_{k=1}^{\mathrm{T}}(g(\bar{\xi}_k)-g(\xi^*)) \leqslant (\frac{2N}{\iota_T}-TN\mu)C_x^2 + C\sum_{k=1}^{\mathrm{T}}\iota_k$, where C is a constant.

Proof. Set $x=\xi^*$. Dividing both sides of (19) by $\frac{2\iota_k}{N}$ and then summing over $k=1,2\ldots T$, we get

$$\sum_{k=1}^{\mathrm{T}}(g(\bar{\xi}_k)-g(\xi^*)) \leqslant \frac{N}{2}\sum_{k=1}^{\mathrm{T}}\frac{1}{\iota_k}\left[\left(1-\frac{\mu\iota_k}{2}\right)\|\bar{\xi}_k-\xi^*\|^2-\|\bar{\xi}_{k+1}-\xi\|^2\right] + \frac{9}{2}NM_d^2\sum_{k=1}^{\mathrm{T}}\iota_k + (2C_g+M_\epsilon)\sum_{k=1}^{\mathrm{T}}\sum_{i=1}^N\|\bar{\xi}_k-\xi_k^i\| + 2M_\epsilon\sum_{k=1}^{\mathrm{T}}\sum_{i=1}^N\|\xi_k^i-\xi^*\| + 2M_d\sum_{i=1}^N\|\bar{\xi}_k-\hat{\xi}_{k+1}^i\|, \tag{20}$$

based on $\|\nabla g^i(\xi)\|\leqslant C_g$ from Remark 5.

Considering

$$\begin{aligned}\sum_{k=1}^{\mathrm{T}}\frac{1}{\iota_k}\left[\left(1-\frac{\mu\iota_k}{2}\right)\|\bar{\xi}_k-\xi^*\|^2-\|\bar{\xi}_{k+1}-\xi\|^2\right] &\leqslant \left(\frac{1}{\iota_1}-\frac{\mu}{2}\right)\|\bar{\xi}_1-\xi^*\|^2-\frac{1}{\iota_T}\|\bar{\xi}_{T+1}-\xi^*\| + \sum_{k=1}^{\mathrm{T}}\left(\frac{1}{\iota_k}-\frac{1}{\iota_{k-1}}-\frac{\mu}{2}\right)\|\bar{\xi}_k-\xi^*\| \\ &\leqslant \left(\frac{1}{\iota_1}-\frac{\mu}{2}\right)4C_x^2 + 4C_x^2\sum_{k=1}^{\mathrm{T}}\left(\frac{1}{\iota_k}-\frac{1}{\iota_{k-1}}-\frac{\mu}{2}\right) \\ &= \left(\frac{1}{\iota_T}-\frac{T\mu}{2}\right)4C_x^2,\end{aligned} \tag{21}$$

it implies $\frac{N}{2}\sum_{k=1}^{\mathrm{T}}\frac{1}{\iota_k}\|\bar{\xi}_k-\xi^*\|^2-\|\bar{\xi}_{k+1}-\xi\|^2] \leqslant (\frac{2N}{\iota_T}-TN\mu)C_x^2$. As for the term $\|\xi_{k+1}^i-\bar{\xi}_{k+1}\|$, by Lemma 4, we obtain

$$\|\xi_{k+1}^i-\bar{\xi}_{k+1}\| \leqslant N\lambda\beta^k\max_{1\leqslant j\leqslant N}\|\xi_0^j\| + \frac{\iota_k}{N}\sum_{j=1}^N\|d_k^j-d_k^i\| + \sum_{s=0}^{k-1}\sum_{j=1}^N\iota_s\lambda\beta^{k-s-1}\|d_s^j\| \leqslant B_1\sum_{s=0}^{k-1}\iota_s\beta^{k-s-1}, \tag{22}$$

where B_1 is a constant. Therefore, $\|\xi_k^i-\bar{\xi}_k\|\leqslant B_1\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1}$ holds. Similarly, $\|\bar{\xi}_k-\hat{\xi}_{k+1}^i\|\leqslant B_2\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1}$ holds for a constant B_2.

Then there is a constant B_3 such as

$$\|\xi_k^i-\xi^*\| = \|\xi_k^i-\bar{\xi}_k\|+\|\bar{\xi}_k-\xi^*\| \leqslant B_1\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1}+2C_x \leqslant B_3\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1}. \tag{23}$$

Combining the above inequalities makes

$$\sum_{k=1}^{\mathrm{T}}(g(\bar{\xi}_k)-g(\xi^*)) \leqslant \frac{4}{\iota_T}C_x^2 + \frac{9NM_d^2}{2}\sum_{k=1}^{\mathrm{T}}\iota_k + C_d\sum_{k=1}^{\mathrm{T}}\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1}, \quad \forall i\in\mathcal{N}, \tag{24}$$

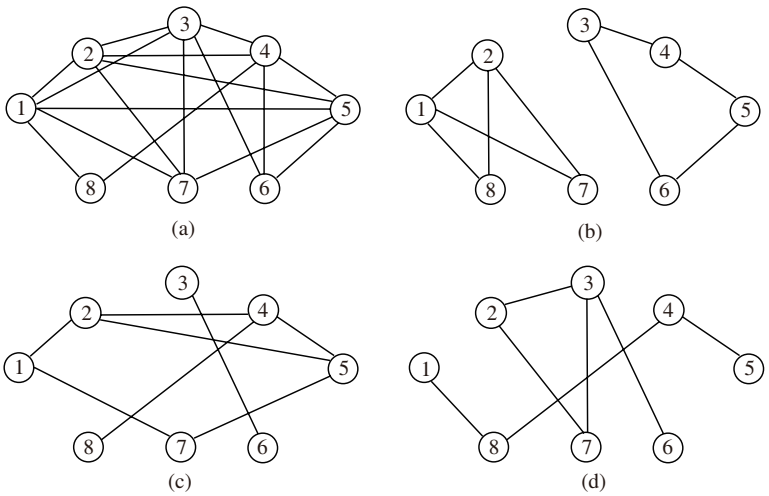

Figure 1 The topology of the networks.

where $C_d = (2C_g + M_\epsilon)NB_1 + 2M_\epsilon NB_3 + 2M_d NB_2$ is a constant. Since

$$\sum_{k=1}^{\mathrm{T}}\sum_{s=1}^{k-1}\iota_{k-s}\beta^{s-1} = \sum_{s=1}^{\mathrm{T}}\beta^{s-1}\sum_{k=s+1}^{\mathrm{T}}\iota_{k-s} \leqslant \sum_{s=1}^{\mathrm{T}}\beta^{s-1}\sum_{k=1}^{\mathrm{T}}\iota_k \leqslant \frac{1}{1-\beta}\sum_{k=1}^{\mathrm{T}}\iota_k,$$

Lemma (4) holds, where $C = C_d\frac{1}{1-\beta} + \frac{9NM_d^2}{2}$ is a constant.

From the above theorem, we can see that the error may increase as the network scale N increases.

Then we consider how to choose the suitable step-size in DAGA for estimation of the regret bound.

Corollary 1. If μ is known, then we can take the step-size $\iota_k = 2/(k\mu)$ for $k > 0$ to achieve $R(T) \backsim O(\log(T)/T)$. If μ is unknown, then we can set $\iota_k = 1/\sqrt{k}$ for $k > 0$ to achieve $R(T) \backsim O(\sqrt{T}/T)$.

Although our algorithm is distributed in missing data cases, the obtained regret analysis is consistent with the online convex optimization [16] and the distributed online learning for privacy-preserving properties [14].

5 Simulations

In this section, two examples are shown to illustrate the performance of Algorithm 1. The first example illustrate the convergence performance of our algorithm, and then the one focuses on the evaluation of the online performance through a time-varying network.

Example 1. Consider an 8-agent network. The communication topology between the agents is shown as Figure 1. Each agent i can obtain a sequence of stationary data $\{A_k^i, y_k^i\}$ for $i = 1, 2, \ldots, 8$, where $\{A_k^i\}$ are i.i.d. normally distributed processes with zero-mean and the covariance matrix $R_k^{A,i} = \mathrm{diag}\{1, 2, 0.5, 0.8, 1.1\}$. The data that each agent i can observe are generated from the linear regression model (4):

$$y_k^i = (A_k^i)^{\mathrm{T}}\xi^i + b_k^i, \quad i = 1, \ldots, 8, \tag{25}$$

where the process noise b_k^i is an i.i.d. normally distributed process with zero-mean and $\sigma_{b_k}^2 = 0.01$. Define the predefined thresholds θ^i for each A_k^i for comparison as follows: θ^i as $[-10, -4, -1, -2, -10]^\top$ for all i and k. Obviously, for the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the area between the interval $[\mu - \sigma, \mu + \sigma]$ is 68.26%, while those between $[\mu - 2\sigma, \mu + 2\sigma]$ and $[\mu - 3\sigma, \mu + 3\sigma]$ are 95.44% and 99.73% respectively [29]. Therefore, data corresponding to A_k^{i1} and A_k^{i5} of vector A_k^i for $i = 1, \ldots, 8$ are lost with relatively small probabilities. Set the unknown optimal vector $\xi^* = [10, 5, 2, 4, 15]^\top$ and use a diminishing step-size $\iota_k = 2.5/k$ for all of the agents. Here the projection set is chosen to be $X = \{\xi \in \mathcal{R}^5 : \|\xi - \xi^*\| \leqslant 80\}$, which is also a bounded convex set. With $\rho_k = 1/k$, then the performance of our algorithm

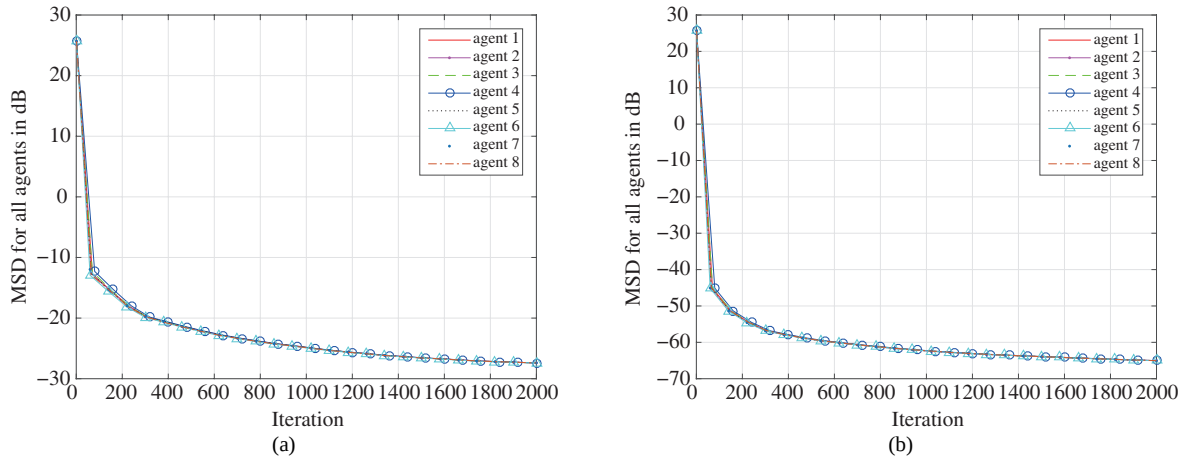

Figure 2 (Color online) The learning curves. (a) Missing data with threshold $[-10,-4,-1,-2,-10]^{\top}$; (b) missing data with threshold $[-10,-10,-10,-10,-10]^{\top}$.

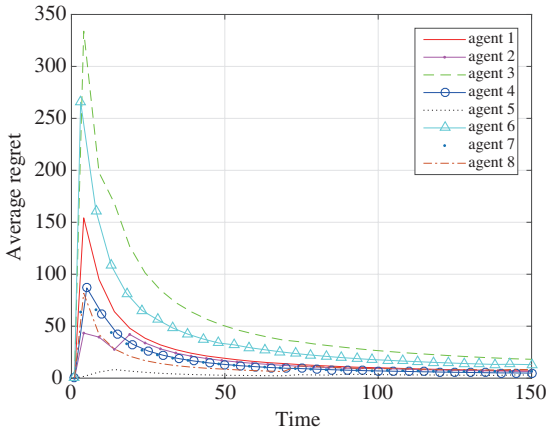

Figure 3 (Color online) The performances of $R(T)$ for all agents.

in 2000 iterate steps can be shown in Figure 2(a). The estimate of ξ^* through 2000 iterate steps is $\hat{\xi} = [9.9874, 5.0001, 1.968, 3.9729, 15.0012]^{\top}$.

If we set the thresholds as $\theta^i = [-10,-10,-10,-10,-10]^{\top}$ for all i,k without changing other settings. It is no surprise that Figure 2(b) shows a much better result than that given in Figure 2(a). The estimate of ξ^* after 2000 iterations is $\hat{\xi} = [10, 5, 2.0004, 3.9999, 15]$.

Example 2. To study the online performance for our algorithm, we suppose that $\{A_k^i, y_k^i\}$ can be generated as before, but each agent can only gradually collect data with respect to time. Still consider an 8-agent network for illustration. The time-varying network topologies between these agents are shown as Figure 1 (b)–(d).

Suppose that the network topology displayed in Figure 1(b) is at time $t = 3k+1,\ k = 0,1,2\ldots$, in Figure 1(c) is at time $t = 3k+2,\ k = 0,1,2,\ldots$, and in Figure 1(d) is at time $t = 3k,\ k = 0,1,2,\ldots$, respectively. Obviously, none of the three graphs are connected, but its union graph is connected and satisfies Assumption 2. According to (18), the regret function during the time interval $[1,T]$ is

$$R(T) = \frac{1}{T}\mathbb{E}\left[\sum_{k=1}^{\mathrm{T}} g(\xi_k) - \sum_{k=1}^{\mathrm{T}} g(\xi^*)\right] = \frac{1}{T}\sum_{k=1}^{\mathrm{T}} \mathbb{E}\left\|y_k^i - (\bar{A}_k^i)^{\mathrm{T}}\xi\right\|^2. \tag{26}$$

Choose the step-size as $\iota_k = 1/\sqrt{k}$ and predefined thresholds as $[-10,-4,-1,-2,-10]^{\top}$ for $i = 1,2,\ldots,8$, with other settings unchanged,

Figure 3 shows the performances of the (average) estimation error $R(T)$ for all agents.

6 Conclusion

In this paper, we discussed the regression estimation with incomplete data in a multi-agent network. To solve the problem we proposed a distributed adaptive gradient-based algorithm with time-varying network topologies and the non-ignorable missing data mechanism. Based on the stochastic approximation technique, we provided a distributed learning algorithm for the non-ignorable missing case and then extended to online case for real-time learning and evaluation. To be strict, we gave the convergence analysis and regret analysis for the algorithms, and moreover, showed various simulation studies to verify the effectiveness of the given distributed estimation algorithm. Note that there are still many challenging problems related to distributed regression design with incomplete data, including distributed analysis using the multiple imputation and improved online learning technology, which are still under our investigation.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2016YFB0901902) and National Natural Science Foundation of China (Grant Nos. 61573344, 61333001, 61374168).

References

1 Nedić A, Ozdaglar A. Distributed subgradient methods for multi-agent optimization. IEEE Trans Automat Control, 2009, 54: 48–61
2 Shi G, Johansson K. Robust consensus for continuous-time multiagent dynamics. SIAM J Control Optim, 2013, 51: 3673–3691
3 Zhang Y Q, Lou Y C, Hong Y G, et al. Distributed projection-based algorithms for source localization in wireless sensor networks. IEEE Trans Wirel Commun, 2015, 43: 3131–3142
4 Feng H, Jiang Z D, Hu B, et al. The incremental subgradient methods on distributed estimations in-network. Sci China Inf Sci, 2014, 57: 092103
5 Lou Y C, Hong Y G, Wang S Y. Distributed continuous-time approximate projection protocols for shortest distance optimization problems. Automatica, 2016, 69: 289–297
6 Yi P, Hong Y G, Liu F. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. Automatica, 2016, 74: 259–269
7 Kokaram A C. On missing data treatment for degraded video and film archives: a survey and a new Bayesian approach. IEEE Trans Image Process, 2004, 13: 397–415
8 Molenberghs G, Kenward M G. Missing Data in Clinical Studies. New York: Wiley, 2007
9 Ibrahim J G, Chen M H, Lipsitz S R, et al. Missing data methods for generalized linear models: a comparative review. J Am Stat Assoc, 2005, 100: 332–346
10 Gholami M R, Jansson M, Strom E G, et al. Diffusion estimation over cooperative multi-agent networks with missing data. IEEE Trans Signal Inf Process Netw, 2016, 2: 276–289
11 Davey A, Savla J. Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach. Oxford, UK: Routledge Academic, 2009
12 Ram S S, Nedić A, Veeravalli V V. Distributed stochastic subgradient projection algorithms for convex optimization. J Optim Theory Appl, 2010, 147: 516–545
13 Graybill F, Iyer H K. Regression Analysis: Concepts and Applications. California: Duxbury Press Belmont, 1994
14 Feng Y, Sundaram S, Vishwanathan S V N, et al. Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties. IEEE Trans Knowl Data Eng, 2013, 25: 2483–2493
15 Hazan E, Kale S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. J Mach Learn Res, 2014, 15: 2489–2512
16 Shamir O, Zhang T. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In: Proceedings of International Conference on Machine Learning, Edinburgh, 2012. 71–79
17 Towfic Z J, Chen J S, Sayed A H. On distributed online classification in the midst of concept drifts. Neurocomputing, 2013, 112: 138–152
18 Widrow B, Stearns S D. Adaptive Signal Processing. Cliffs: Prentice-Hall, 1985. 1–32
19 Sayed A H. Adaptation, learning, and optimization over networks. Found Trends Mach Learn, 2014, 7: 311–801
20 Sayed A H, Tu S Y, Chen J S, et al. Diffusion strategies for adaptation and learning. IEEE Signal Proc Mag, 2013, 30: 155–171
21 Polyak B T. Introduction to Optimization. New York: Optimization Software Inc., 1983. 2–8
22 Godsil C, Royle G. Algebraic Graph Theory. New York: Springer-Verlag, 2001. 1–18
23 Ferguson T S. A Course in Large Sample Theory. London: Chapman and Hall Ltd., 1996. 3–4
24 Durrett R. Probability Theory and Examples. Camberidge, UK: Camberidge Press, 2010. 328–347
25 Enders C K. Applied Missing Data Analysis. New York: The Guilford Press, 2010

26 Kushner H J, Yin G. Stochastic Approximation and Recursive Algorithms and Applications. New York: Springer-Verlag, 1997. 117–157
27 Widrow B, Mccool J, Larimore M G, et al. Stationary and nonstationary learning characteristics of the LMS adaptive filter. Proc IEEE, 1976, 64: 1151–1162
28 Yi P, Hong Y G. Stochastic sub-gradient algorithm for distributed optimization with random sleep scheme. Control Theory Technol, 2015, 13: 333–347
29 Larsen R J, Max M L. An Introduction to Mathematical Statistics and Its Applications. 4th ed. New York: Pearson, 2006. 221–280

Appendix A Proof of Lemma 3

With the observation noise of (13), we obtain

$$\mathbb{E}\|\epsilon_k^i\| = \mathbb{E}\|R_k^{\bar{A},i}\xi_k^i - R^{\bar{A},i}\xi_k^i + r^{y\bar{A},i} - y_k^i\bar{A}_k^i\| \leqslant \mathbb{E}\|R_k^{\bar{A},i} - R^{\bar{A},i}\|\|\xi_k^i\| + \mathbb{E}\|r^{y\bar{A},i} - y_k^i\bar{A}_k^i\|, \quad \forall i \in \mathcal{N}. \tag{A1}$$

Therefore, for all $\epsilon > 0$, there exists a k_1 (k_1 is an integer) for $k > k_1$, such that $\|R_k^{\bar{A},i} - R^{\bar{A},i}\| < \epsilon$. Define $M_1 = \max\{\|R_1^{\bar{A},i} - R^{\bar{A},i}\|, \|R_2^{\bar{A},i} - R^{\bar{A},i}\|, \ldots, \|R_{k_1}^{\bar{A},i} - R^{\bar{A},i}\|, \epsilon\}$. Then $\|R_k^{\bar{A},i} - R^{\bar{A},i}\| \leqslant M_1$ for $k \geqslant 0$. Analogously, $\|r_k^{y\bar{A},i} - R^{\bar{A},i}\| \leqslant M_2$ for $k \geqslant 0$. From Remark 3, we have $\|\xi_k^i\| < C_x$. Hence, $\mathbb{E}\|\epsilon_i(k)\| \leqslant M_1C_x + M_2 = M_\epsilon, \forall k \geqslant 0$. By (13), $d_k^i = \nabla g_i(k) + \epsilon_i(k)$. Thus, $\mathbb{E}\|d_k^i\| \leqslant \mathbb{E}\|\nabla g_i(k)\| + \mathbb{E}\|\epsilon_i(k)\| \leqslant C_g + M_\epsilon = M_d$, which is bounded.

Appendix B Proof of Lemma 4

For all $i \in \mathcal{N}$, $k \geqslant 0$, define $p_{k+1}^i = \xi_{k+1}^i - \sum_{j=1}^N w_{ij}(k)\xi_k^j$. We rewrite (8) compactly in terms of $\Psi(k,s)$ as follows: $\xi_{k+1}^i = \sum_{j=1}^N[\Psi(k,0)]_{ij}\xi_0^j + p_{k+1}^i + \sum_{s=1}^k\sum_{j=1}^N[\Psi(k,s)]_{ij}p_s^j$, for $k \geqslant s$. Moreover, with Assumption 1 and by induction, the following equality holds: $\bar{\xi}_{k+1} = \frac{1}{N}\sum_{i=1}^N \xi_0^i + \frac{1}{N}\sum_{s=1}^{k+1}\sum_{j=1}^N p_s^j$. Consequently, we obtain that, for $i \in \mathcal{N}$, $\xi_{k+1}^i - \bar{\xi}_{k+1} = \sum_{j=1}^N([\Psi(k,0)]_{ij} - \frac{1}{N})\xi_0^j + (p_{k+1}^i - \frac{1}{N}\sum_{j=1}^N p_{k+1}^j) + \sum_{s=1}^k\sum_{j=1}^N([\Psi(k,0)]_{ij} - \frac{1}{N})p_s^j$. Therefore, $\forall i \in \mathcal{N}$,

$$\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| \leqslant \sum_{j=1}^N|[\Psi(k,0)]_{ij} - \frac{1}{N}|\|\xi_0^i\| + \|p_{k+1}^i\| + \left\|\frac{1}{N}\sum_{j=1}^N p_{k+1}^j\right\| + \sum_{s=1}^k\sum_{j=1}^N|[\Psi(k,0)]_{ij} - \frac{1}{N}|\|p_s^j\|. \tag{B1}$$

Plugging in the estimate of $\Psi(k,s)$ in Lemma 1 and $\|\xi_0^i\| \leqslant \max_{1\leqslant i\leqslant N}\|\xi_0^i\|$, we have

$$\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| \leqslant N\lambda\beta^k \max_{1\leqslant i\leqslant N}\|\xi_0^i\| + \|p_{k+1}^i\| + \frac{1}{N}\sum_{j=1}^N\|p_{k+1}^j\| + \lambda\sum_{s=1}^k\beta^{k-s}\sum_{j=1}^N\|p_s^j\|. \tag{B2}$$

Next, from the definition of $p_i(k)$, we get

$$\|p_{k+1}^i\| = \left\|P_X\left(\sum_{j=1}^N w_{ij}(k)\xi_k^i - \iota_k d_k^i\right) - \sum_{j=1}^N w_{ij}(k)\xi_k^j\right\| \leqslant \iota_k\|d_k^i\|. \tag{B3}$$

With (B2) and (B3), the proof is completed.

Appendix C Proof of Theorem 2

From Theorem 1, $\|\xi_{k+1}^i - \bar{\xi}_{k+1}\|$ converges in mean. Then, on the base of Fatou's Lemma[1)], the following relation holds $0 \leqslant \mathbb{E}[\liminf_{k\to\infty}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\|] \leqslant \liminf_{k\to\infty}\mathbb{E}[\|\xi_{k+1}^i - \bar{\xi}_{k+1}\|] = 0$, which yields $\mathbb{E}[\liminf_{k\to\infty}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\|] = 0$. Therefore,$\liminf_{k\to\infty}\|\xi_{k+1}^i - \bar{\xi}_{k+1}\| = 0$ holds almost surely. Since $\|\xi_{k+1}^i - \bar{\xi}_k\|^2 \leqslant \|\hat{\xi}_{k+1}^i - \bar{\xi}_k\|^2$,

$$\|\xi_{k+1}^i - \bar{\xi}_k\|^2 \leqslant \|\hat{\xi}_{k+1}^i - \bar{\xi}_k\|^2 \leqslant \sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|^2 + \iota_k^2\|d_k^i\|^2 + 2\iota_k\|d_k^i\|\sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|. \tag{C1}$$

Note that $\sum_{i=1}^N\|\xi_{k+1}^i - \bar{\xi}_k\|^2 \leqslant \sum_{i=1}^N\sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|^2 + \sum_{i=1}^N\iota_k^2\|d_k^i\|^2 + 2\sum_{i=1}^N\iota_k\|d_k^i\|\sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|$, $i \in \mathcal{N}$, which implies $\sum_{i=1}^N\sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|^2 = \sum_{i=1}^N\|\xi_k^i - \bar{\xi}_k\|^2$. Therefore,

$$\sum_{i=1}^N\|\xi_{k+1}^i - \bar{\xi}_k\|^2 \leqslant \sum_{i=1}^N\|\xi_k^i - \bar{\xi}_k\|^2 + \sum_{i=1}^N\iota_k^2\|d_k^i\|^2 + 2\sum_{i=1}^N\iota_k\|d_k^i\|\sum_{j=1}^N w_{ij}(k)\|\xi_k^j - \bar{\xi}_k\|. \tag{C2}$$

Taking the conditional expectation of both side of (C2) yields

$$\sum_{i=1}^N\mathbb{E}[\|\xi_{k+1}^i - \bar{\xi}_{k+1}\|^2|F_k] \leqslant \sum_{i=1}^N\|\xi_k^i - \bar{\xi}_k\|^2 + 2M_d\sum_{j=1}^N\iota_k\|\xi_k^j - \bar{\xi}_k\| + N\iota_k^2M_d^2. \tag{C3}$$

According to Theorem 6.2 of [12], $\sum_{k=1}^\infty\iota_k\|\xi_k^j - \bar{\xi}_k\| < \infty$ with probability 1. Therefore, together with $\sum_{k=1}^\infty N\iota_k^2M_d^2 < \infty$, $\|\xi_{k+1}^i - \bar{\xi}_k\|^2$ converges almost surely by Lemma 2. Hence, the conclusion follows.

1) Rudin W. Real and Complex Analysis. New York: McGraw-Hill Book Company, 1986. 5–71.

Appendix D Proof of Theorem 3

Clearly, $\|\xi_{k+1}^i-\xi^*\|^2 \leqslant \|\hat{\xi}_{k+1}^i-\xi^*\|^2$, and then $\|\xi_{k+1}^i-\xi^*\|^2 \leqslant \|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\|^2+\iota_k^2\|d_k^i\|^2-2\iota_k(d_k^i)^{\mathrm{T}}$ $(\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*)$, which follows from [21] that, $\forall x_1, x_2$, $g(x_2) \geqslant g(x_1)+\nabla g(x_1)^{\mathrm{T}}(x_2-x_1)$. Recalling that, $\mathbb{E}\|d_k^i\| \leqslant M_d$ in Lemma 3 and $\|\nabla g^i(\xi)\| \leqslant C_g$ in Remark 5, we have

$$(\nabla g_k^i)^{\mathrm{T}}\left(\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right) \geqslant g^i(\bar{\xi}_k)-g_i(\xi^*)-C_g\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\bar{\xi}_k\right\|, \tag{D1}$$

and $\mathbb{E}[\epsilon_i^{\mathrm{T}}(k)(\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*)] \leqslant \mathbb{E}\|\epsilon_k^i\|\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\|$ for all $k=0,1,2,\ldots$. Therefore,

$$\begin{aligned}\mathbb{E}[\|\xi_{k+1}^i-\xi^*\|^2|F_k] \leqslant{}& \mathbb{E}\left[\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right\|^2\Big|F_k\right]+\iota_k^2\mathbb{E}\|d_k^i\|^2+2\iota_k C_g\mathbb{E}\left[\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\bar{\xi}_k\right\|\Big|F_k\right]\\ &-2\iota_k\mathbb{E}\|\epsilon_k^i\|\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right\|-2\iota_k(g_i(\bar{\xi}_k)-g_i(\xi^*)).\end{aligned} \tag{D2}$$

By the double stochasticity of matrix $W(k)$,

$$\begin{cases}\displaystyle\sum_{i=1}^n\mathbb{E}\left[\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right\|^2\Big|F_k\right] \leqslant \sum_{i=1}^n\|\xi_k^i-\xi^*\|^2,\\ \displaystyle\sum_{i=1}^n\mathbb{E}\left[\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\bar{\xi}_k\right\|\Big|F_k\right] \leqslant \sum_{i=1}^n\|\xi_k^i-\bar{\xi}_k\|.\end{cases} \tag{D3}$$

Then, with probability 1, for $i\in\mathcal{N}$, it holds

$$\sum_{i=1}^N\mathbb{E}[\|\xi_{k+1}^i-\xi^*\|^2|F_k] \leqslant \sum_{i=1}^N\|\xi_k^i-\xi^*\|^2+w_k-v_k, \tag{D4}$$

where

$$\begin{cases}w_k=\displaystyle\sum_{i=1}^N\iota_k^2\mathbb{E}\|d_k^i\|^2+2\iota_k C_g\sum_{i=1}^N\mathbb{E}\|\xi_k^i-\bar{\xi}_k\|,\\ v_k=2\displaystyle\sum_{i=1}^N\iota_k\mathbb{E}[\|\epsilon_k^i\|]\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right\|+2\iota_k g(\bar{\xi}_k-g(\xi^*)).\end{cases} \tag{D5}$$

By Theorem 6.2 in [12], $\sum_{k=1}^\infty 2\iota_k C_g\sum_{i=1}^N\mathbb{E}\|\xi_k^i-\bar{\xi}_k\|<\infty$. Since $\sum_{k=1}^\infty\iota_k^2<\infty$, $\sum_{k=1}^\infty\iota_k^2\mathbb{E}\|d_k^i\|^2 \leqslant \sum_{k=1}^\infty\iota_k^2 NM_d^2<\infty$. Therefore, $\sum_{k=1}^\infty w_k<\infty$.

From Lemma 2, the sequence $\sum_{i=1}^N\|\xi_k^i-\xi^*\|^2$ converges with probability 1 and $\sum_{k=1}^\infty v_k<\infty$.

As for v_k, according to the boundedness of ξ_k^i and the ergodicity of $\bar{A}_k^i$, we conclude that $\lim_{k\to\infty}R_k^{\bar{A},i}\xi_k^i-R^{\bar{A},i}\xi_k^i=0$. Moreover, $\lim_{k\to\infty}r^{y\bar{A},i}-y_k^i\bar{A}_k^i=0$ by the stationary property of y_k^i, a_k^i.

Therefore

$$\lim_{k\to\infty}2\sum_{i=1}^N\mathbb{E}[\|\epsilon_k^i\|]\left\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\right\|=0. \tag{D6}$$

Similar to the demonstration of Theorem 6.2 in [12], we get $\sum_{k=1}^\infty 2\sum_{i=1}^N\iota_k\mathbb{E}[\|\epsilon_k^i\|]\|\sum_{j=1}^N w_{ij}(k)\xi_k^i-\xi^*\|<\infty$, which implies $\sum_{k=1}^\infty 2\iota_k(g(\bar{\xi}_k)-g(\xi^*))<\infty$. Since $\sum_{k=1}^\infty 2\iota_k(g(\bar{\xi}_k)-g(\xi^*))<\infty$ and $\sum_{i=1}^\infty\iota_k=\infty$, $\liminf_{k\to\infty}g(\bar{\xi}_k)=g(\xi^*)$ holds almost surely. Therefore, $\lim_{k\to\infty}\|\xi_k^i-\bar{\xi}_k\|=0$ holds almost surely for all i, which yields the conclusion.

Appendix E Proof of Lemma 5

Define $r_k^i=\xi_k^i-\hat{\xi}_k^i=P_X(\hat{\xi}_k^i)-\hat{\xi}_k^i$. Since X is convex and $W(k)$ is doubly stochastic, we have $\sum_{j=1}^n w_{ij}(k)x_k^j\in X$, which leads to $\|r_{k+1}^i\|\| \leqslant \|P_X(\hat{\xi}_{k+1}^i)-\sum_{j=1}^n w_{ij}(k)\xi_k^j\|+\iota_k\|d_k^i\| \leqslant 2\iota_k\|d_k^i\|$. By Algorithm 1, we obtain $\bar{\xi}_{k+1}=\bar{\xi}_k-\frac{\iota_k}{N}\sum_{i=1}^N(\nabla g_k^i+\epsilon_k^i)+\frac{1}{N}\sum_{i=1}^N r_{k+1}^i$. As a result, we can decompose $\|\bar{\xi}_{k+1}-\xi\|^2$ by

$$\|\bar{\xi}_{k+1}-\xi\|^2=\|\bar{\xi}_k-\xi\|^2+\frac{1}{N^2}\left\|\sum_{i=1}^N(r_{k+1}^i+\iota_k d_k^i)\right\|^2+\frac{2}{N}\sum_{i=1}^N\langle r_{k+1}^i,\bar{\xi}_k-\xi\rangle-\frac{2\iota_k}{N}\sum_{i=1}^N\langle\nabla g_k^i,\bar{\xi}_k-\xi\rangle-\frac{2\iota_k}{N}\sum_{i=1}^N\langle\epsilon_k^i,\bar{\xi}_k-\xi\rangle. \tag{E1}$$

Let us check $-\sum_{i=1}^N\langle\nabla g_k^i,\bar{\xi}_k-\xi\rangle$. Based on Lemma (1), we obtain

$$\begin{aligned}-\langle\nabla g_k^i,\bar{\xi}_k-\xi\rangle={}&-\langle\nabla g_k^i,\bar{\xi}_k-\xi_k^i\rangle-\langle\nabla g_k^i,\xi_k^i-\xi\rangle \leqslant \|\nabla g_k^i\|\|\bar{\xi}_k-\xi_k^i\|+g^i(\bar{\xi}_k)-g_k^i-\frac{\mu}{2}\|\xi_k^i-x\|^2+g^i(\xi)-g^i(\bar{\xi}_k)\\ \leqslant{}&\|\nabla g_k^i\|\|\bar{\xi}_k-\xi_k^i\|+\langle\nabla\bar{g}_k^i,\bar{\xi}_k-\xi_k^i\rangle-\frac{\mu}{2}\|\xi_k^i-\xi\|^2-\frac{\mu}{2}\|\xi_k^i-\bar{\xi}_k\|^2+g^i(\xi)-g^i(\bar{\xi}_k).\end{aligned} \tag{E2}$$

Since $\langle \nabla \bar{g}_k^i, \bar{\xi}_k - \xi_k^i \rangle \leqslant \|\nabla \bar{g}_k^i\| \|\bar{\xi}_k - \xi_k^i\|$ and $\|\xi_k^i - \xi\|^2 + \|\xi_k^i - \bar{\xi}_k\|^2 \geqslant \frac{1}{2}\|\bar{\xi}_k - \xi\|^2$, we can estimate $-\langle \nabla g_k^i, \bar{\xi}_k - \xi \rangle$ as follows: $-\langle \nabla g_k^i, \bar{\xi}_k - \xi \rangle \leqslant (\|\nabla g_k^i\| + \|\nabla \bar{g}_k^i\|)\|\bar{\xi}_k - \xi_k^i\| + g^i(\xi) - g^i(\bar{\xi}_k) - \frac{\mu}{4}\|\bar{\xi}_k - \xi\|^2$.

Summing up over $i = 1, 2, \ldots, N$, the following inequality holds:

$$-\sum_{i=1}^{N} \langle \nabla g_k^i, \bar{\xi}_k - \xi \rangle \leqslant \sum_{i=1}^{N} (\|\nabla g_k^i\| + \|\nabla \bar{g}_k^i\|)\|\bar{\xi}_k - \xi_k^i\| + g(\xi) - g(\bar{\xi}_k) - \frac{\mu N}{4}\|\bar{\xi}_k - \xi\|^2. \tag{E3}$$

Next, it is not hard to get that, for $k = 0, 1, \ldots$,

$$-\sum_{i=1}^{N} \langle \epsilon_k^i, \bar{\xi}_k - \xi \rangle \leqslant \sum_{i=1}^{N} \|\epsilon_k^i\| \|\bar{\xi}_k - \xi_k^i\| + \sum_{i=1}^{N} \|\epsilon_k^i\| \|\xi_k^i - \xi\|. \tag{E4}$$

Then

$$\langle r_{k+1}^i, \bar{\xi}_k - \xi \rangle \leqslant \langle r_{k+1}^i, \bar{\xi}_k - \hat{\xi}_{k+1}^i \rangle + \langle P_X(\hat{\xi}_{k+1}^i) - \hat{\xi}_{k+1}^i, \hat{\xi}_{k+1}^i - \xi \rangle. \tag{E5}$$

Because the projection operator satisfies the following inequality

$$\langle P_X(\hat{\xi}) - \hat{\xi}, \hat{\xi} - \xi \rangle \leqslant -\|P_X(\hat{\xi}) - \hat{\xi}\|^2 \leqslant 0, \ \forall \xi \in X, \tag{E6}$$

it follows from (E6) with (E5) that

$$\langle r_{k+1}^i, \bar{\xi}_k - \xi \rangle \leqslant \langle r_{k+1}^i, \bar{\xi}_k - \hat{\xi}_{k+1}^i \rangle \leqslant 2\iota_k \|d_k^i\| \|\bar{\xi}_k - \hat{\xi}_{k+1}^i\|. \tag{E7}$$

Moreover,

$$\frac{1}{N^2} \left\| \sum_{i=1}^{N} (r_{k+1}^i + \iota_k d_k^i) \right\|^2 = \frac{1}{N^2} \left(\sum_{i=1}^{N} (r_{k+1}^i + \iota_k d_k^i) \right)^2 \leqslant \frac{9\iota_k^2}{N^2} \left(\sum_{i=1}^{N} \|d_k^i\| \right)^2. \tag{E8}$$

Combining (E3), (E4), (E7), and (E8) with (E1) yields the conclusion.