SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Distributed gradient-based sampling algorithm for least-squares in switching multi-agent networks

Peng Lin[1,2] & Hongsheng Qi[1,2*]

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

Appendix A Distributed Gradient-based Sampling Algorithm

Algorithm A1 Distributed Gradient-based Sampling Algorithm (DGSA)

Input: The datasets: $\mathcal{D} = \{D_i\}_{i=1}^N$, the pilot estimate ω_0, N and M_s;
Output: Estimations $\{\hat{\omega}_i\}_{i=1}^N$ of ω^*;
The Sampling Step:
for $i = 1, 2 \cdots, N$ **do**
  for $k = 1, 2 \cdots, m_i$ **do**
    Calculate $||g_i^k|| = ||x_i^k(y_i^k - \omega_0^\top x_i^k)||$;
  end for
  Calculate $z_i(0) = \sum_{k=1}^{m_i} ||g_i^k||$;
end for
for $t = 1, 2, \cdots$ **do**
  for $i = 1, 2 \cdots, N$ **do**
    Calculate $z_i(t+1) = \sum_{j \in \mathcal{N}_i(t)} a_{i,j}(t) z_j(t)$ until consensus;
  end for
end for
for $i = 1, 2, \cdots, N$ **do**
  for $k = 1, 2, \cdots, m_i$ **do**
    Calculate sampling probability $\beta_i^k = \frac{||g_i^k||}{N z_i(T)}$;
    Generate $s_i^k \sim$ Bernoulli$(1, p_i^k)$;
    if $s_i^k = 1$ **then**
      This data point is chosen;
    end if
  end for
end for
The Optimization Step:
for $t = 1, 2, \cdots$ **do**
  for $i = 1, 2 \cdots, N$ **do**
    Calculate $\omega_i(t+1) = P_\Gamma[\sum_{j=1}^N a_{i,j}(t)\omega_j(t) - \alpha(t)d_i(t)]$ until convergence;
  end for
end for

Remark 1. In DGSA, it needs the pilot estimate ω_0 of ω^*. It can be obtained from an estimation from an initial subsample of size M_{s_0} with uniform sampling. Therefore, all agents need to communicate with their one-hop neighbors to achieve the total number of the data points in the network. The same idea like the sampling step, let $z_i(0) = m_i$, and

* Corresponding author (email: qihongsh@amss.ac.cn)

all agents calculate $z_i(t+1)$ until consensus. The uniform sampling probability is $\beta_i^k = (Nz_i(T))^{-1}$. After the subsample dataset $\mathcal{D}_{s_0} = \{D_i^{s_0}\}_{i=1}^N$ is determined, all agents do the optimization step to minimize the following equation to get ω_0:

$$F_{s_0}(X_{s_0}, Y_{s_0}, \omega) = \sum_{i=1}^{N} f_i(X_i^{s_0}, Y_i^{s_0}, \omega),$$

where $f_i(X_i^{s_0}, Y_i^{s_0}, \omega) = \sum_{k\in D_i^{s_0}} (y_i^k - \omega^\top x_i^k)^2$.

Remark 2. The actual size M_a of sampling dataset is a random variable because of the Poisson sampling, with $\mathbb{E}[M_a] = M_s$. Therefore, the expected sample size is initialized as an input. In DGSA, some β_i^k may be so large that $p_i^k > 1$. In this circumstance, we just let $p_i^k = 1$ and choose this data point.

Remark 3. The communication network in DGSA are all assumed to be time-varying to be applicable to failures of the links among the agents and the reduction of the consumption of communication, which is more robust and safe.

Appendix B Proof of Theorem 1

The following lemma is given to help our analysis of DGSA.

Lemma 1 (see [1,2])**.** Consider a finite Hermitian matrices $\{B_i^k = x_i^k x_i^{k\top}\}$. Let $\{\gamma_i^k\}$ be a finite sequence of independent Bernoulli variables, whose mean is p_i^k, respectively. Let $R_x = \max_{i,k} ||x_i^k||^2$ and $R_\Sigma = M^{-2}\sum_{i=1}^N \sum_{k=1}^{m_i} (\beta_i^k)^{-1} ||x_i^k||^4$. The random matrix Δ is defined as

$$\Delta = M^{-1} \sum_{i=1}^{N} \sum_{k=1}^{m_i} (1 - \gamma_i^k / p_i^k) B_i^k.$$

Then we have

$$\mathbb{E}[\lambda_{\max}(\Delta)] \leqslant M_s^{-1/2} R_\Sigma \sqrt{2\log d} + R_x \log d / 3M,$$

where d is the dimension of vector x_i^k.

Theorem 1. Denote $R_b^2 = M^{-2}\sum_{i=1}^N \sum_{k=1}^{m_i} (\beta_i^k)^{-1} ||\epsilon_i^k x_i^k||^2$ where $\epsilon_i^k = y_i^k - \omega^{*\top} x_i^k$, and if M_s satisfies

$$M_s > 2R_\Sigma^2 \log d / (2^{-1}\delta\lambda_{min}(\Omega_M) - (3M)^{-1} R_x \log d)^2$$

for any $\delta > 2R_x \log d / 3M\lambda_{min}(\Omega_M)$, it is obtained that

$$\mathbb{P}(||\tilde{\omega} - \omega^*|| \leqslant C_M M_s^{-1/2}) \geqslant 1 - \delta,$$

where $C_M = 3\lambda_{min}(\Omega_M)\delta^{-1} R_b$.

Proof. According to the definition of ω^* and $\tilde{\omega}$, we obtain

$$||\tilde{\omega} - \omega^*|| = ||\Omega_s^{-1} b_s - \Omega_s^{-1}\Omega_s \omega^*|| \leqslant \lambda_{max}(\Omega_s^{-1}) ||b_s - \Omega_s \omega^*||. \tag{B1}$$

It is obvious that we have the following relations:

$$\lambda_{max}(\Omega_s^{-1}) - \lambda_{max}(\Omega_M^{-1}) \leqslant \lambda_{max}(\Omega_s^{-1} - \Omega_M^{-1}) \leqslant \lambda_{max}(\Omega_s^{-1})\lambda_{max}(\Omega_M^{-1})\lambda_{max}(\Omega_M - \Omega_s). \tag{B2}$$

Then we obtain

$$\lambda_{max}(\Omega_s^{-1}) \leqslant \lambda_{max}(\Omega_M^{-1}) / (1 - \lambda_{max}(\Omega_M^{-1})\lambda_{max}(\Omega_M - \Omega_s)). \tag{B3}$$

If the following event holds

$$\mathcal{E}_1 = \{\lambda_{max}(\Omega_M - \Omega_s) < \lambda_{min}(\Omega_M)/2\}, \tag{B4}$$

then (B1) can be transformed into

$$||\tilde{\omega} - \omega^*|| \leqslant (\lambda_{min}^{-1}(\Omega_M) + 2\lambda_{min}^{-2}(\Omega_M)\lambda_{max}(\Omega_M - \Omega_s)) ||b_s - \Omega_s \omega^*||. \tag{B5}$$

For any $\delta > 0$, denote the event $\mathcal{E}_2$ as

$$\mathcal{E}_2 = \{||b_s - \Omega_s \omega^*|| \leqslant R_b / \delta\sqrt{M_s}\}, \tag{B6}$$

Because $\mathbb{E}[||b_s - \Omega_s\omega^*||^2] < M^{-2}\sum_{i=1}^N \sum_{k=1}^{m_i} ||\epsilon_i^k x_i^k||^2$, we obtain $\mathbb{P}(\bar{\mathcal{E}_2}) \leqslant \delta$ with Markov's inequality. The event $\mathcal{E}_3$ is defined as

$$\mathcal{E}_3 = \{\lambda_{max}(\Omega_M - \Omega_s) \leqslant R_\Sigma \sqrt{2\log d} / \sqrt{M_s}\delta + R_x \log d / 3M\delta\}. \tag{B7}$$

By combining with Lemma 1, we obtain $\mathbb{P}(\bar{\mathcal{E}_3}) \leqslant \delta$. The relation $\bar{\mathcal{E}_1} \subset \bar{\mathcal{E}_3}$ can be obtained if the following relations hold

$$\begin{aligned} M_s &> 2R_\Sigma^2 \log d / (2^{-1}\delta\lambda_{min}(\Omega_M) - (3M)^{-1} R_x \log d)^2, \\ \delta &> 2R_x \log d / 3M\lambda_{min}(\Omega_M). \end{aligned} \tag{B8}$$

Thus, we obtain

$$\mathbb{P}(||\tilde{\omega} - \omega^*|| \leqslant C_M M_s^{-1/2}) \geqslant 1 - \delta, \tag{B9}$$

where $C_M = 3\lambda_{min}(\Omega_M)\delta^{-1} R_b$.

Appendix C Proof of Theorem 2

For the convenience of the proof, several lemmas are given first. Lemma 2 is always used in the proof with time-varying topology.

Lemma 2 (see [3]). Under Assumption 1, for all i, j and all $t_1 \leqslant t_2$, we have

$$\left|[\varphi(t_1:t_2)]_{i,j} - \frac{1}{N}\right| \leqslant \zeta^{-2}\varrho^{t_2-t_1+1},$$

where $\zeta = 1 - \frac{\epsilon}{4N^2}$, $\varrho = \zeta^{1/\tau}$, $\varphi(t_1 : t_2)$ is a transition matrix defined by $\varphi(t_1 : t_2) = A(t_1)A(t_1+1)\cdots A(t_2)$ with $\varphi(t_1 : t_1) = A(t_1)$.

Lemma 3 gives the relation between the estimate $\omega_i(t+1)$ obtained by agent i at time $t+1$ and any point ω in the common domain Γ, whose proof is in Appendix C.1.

Lemma 3. With Assumption 1 holds, the following relation can be obtained

$$\sum_{i=1}^{N}||\omega_i(t+1)-\vartheta||^2 \leqslant \sum_{j=1}^{N}||\omega_j(t)-\vartheta||^2 + \alpha^2(t)\sum_{i=1}^{N}||d_i(t)||^2 - 2\alpha(t)\sum_{i=1}^{N}(\tilde{f}_i(v_i(t)) - \tilde{f}_i(\vartheta)) - \sum_{i=1}^{N}||\phi_i(t)||^2.$$

Denote $\bar{\omega}(t) = \frac{1}{N}\sum_{i=1}^{N}\omega_i(t)$. Lemma 4 shows all agents will arrive at the average of the states, whose proof is given in Appendix C.2.

Lemma 4. With Assumption 1 and the step condition hold, we have $||\omega_i(t)-\bar{\omega}(t)|| \to 0$, as $t \to \infty$, for all $i = 1, 2, \ldots, N$.

Lemma 5 shows that the accumulated errors multiplied by the step-size for all agents is bounded, whose proof is in Appendix C.3.

Lemma 5. Let Assumption 1 and step-size condition hold, we have $\sum_{t=1}^{\infty}\alpha(t)||\omega_i(t)-\bar{\omega}(t)|| < \infty$ for all $i = 1, 2, \ldots, N$.

Proof. According to Lemma 3, we obtain

$$\begin{aligned}\sum_{i=1}^{N}||\omega_i(t+1)-\vartheta||^2 \leqslant& \sum_{j=1}^{N}||\omega_j(t)-\vartheta||^2 + NL^2\alpha^2(t) - 2\alpha(t)\sum_{i=1}^{N}(\tilde{f}_i(v_i(t)) - \tilde{f}_i(\bar{\omega}(t)))\\ &- 2\alpha(t)\sum_{i=1}^{N}(\tilde{f}_i(\bar{\omega}(t)) - \tilde{f}_i(\vartheta)).\end{aligned} \tag{C1}$$

Because of the subgradient properties and the convexity of $||\cdot||$, we obtain

$$|\tilde{f}_i(v_i(t)) - \tilde{f}_i(\bar{\omega}(t))| \leqslant L||v_i(t)-\bar{\omega}(t)|| \leqslant L\sum_{j=1}^{N}a_{i,j}(t)||\omega_j(t)-\bar{\omega}(t)||. \tag{C2}$$

Hence (C1) can be transformed into

$$\begin{aligned}\sum_{i=1}^{N}||\omega_i(t+1)-\vartheta||^2 \leqslant& \sum_{i=1}^{N}||\omega_i(t)-\vartheta||^2 + NL^2\alpha^2(t) + 2L\alpha(t)\sum_{j=1}^{N}||\omega_j(t)-\bar{\omega}(t)||\\ &- 2\alpha(t)(\tilde{F}(\bar{\omega}(t)) - \tilde{F}(\vartheta)).\end{aligned} \tag{C3}$$

Let $\vartheta = \tilde{\omega}$ and sum (C3) from $t = T_1$ to T_2 with $T_2 > T_1$, and we obtain

$$\begin{aligned}\sum_{i=1}^{N}||\omega_i(T_2+1)-\tilde{\omega}||^2 + \sum_{t=T_1}^{T_2}2\alpha(t)(\tilde{F}(\bar{\omega}(t)) - \tilde{F}(\tilde{\omega})) \leqslant& \sum_{i=1}^{N}||\omega_i(T_1)-\tilde{\omega}||^2 + NL^2\sum_{T_1}^{T_2}\alpha^2(t)\\ &+ 2L\sum_{t=T_1}^{T_2}\alpha(t)\sum_{j=1}^{N}||\omega_j(t)-\bar{\omega}(t)||.\end{aligned} \tag{C4}$$

Let $T_1 = 1$, $T_2 \to \infty$, and according to the step-condition and Lemma 3, we claim that

$$\sum_{t=1}^{\infty}\alpha(t)(\tilde{F}(\bar{\omega}(t)) - \tilde{F}(\tilde{\omega})) < \infty. \tag{C5}$$

Combining (C5) with $\sum_{t=1}^{\infty}\alpha(t) = \infty$, it implies

$$\liminf_{t\to\infty}(\tilde{F}(\bar{\omega}(t)) - \tilde{F}(\tilde{\omega})) = 0. \tag{C6}$$

Because $\tilde{\omega}$ is the optimal point of $\tilde{F}$, $\tilde{F}(\bar{\omega}(t)) - \tilde{F}(\tilde{\omega})$ is nonnegative. Hence we have

$$\sum_{i=1}^{N}||\omega_i(T_2+1)-\tilde{\omega}||^2 \leqslant \sum_{i=1}^{N}||\omega_i(T_1)-\tilde{\omega}||^2 + NL^2\sum_{T_1}^{T_2}\alpha^2(t) + 2L\sum_{t=T_1}^{T_2}\alpha(t)\sum_{j=1}^{N}||\omega_j(t)-\bar{\omega}(t)||. \tag{C7}$$

Therefore, we obtain that $\{\omega_i(t)\}$ is bounded and $\{\sum_{i=1}^{N}||\omega_i(t)-\tilde{\omega}||\}$ is convergent. According to Lemma 4, we know that $\{||\bar{\omega}(t)-\tilde{\omega}||\}$ is convergent and $\{\bar{\omega}(t)\}$ have a limit point. From (C6) and the continuity of $\tilde{F}$, we obtain that $\lim_{t\to\infty}\bar{\omega}(t) = \tilde{\omega}$. Combining with Lemma 4, it implies that $\omega_i(t)$ converges to $\tilde{\omega}$ for all $i = 1, 2, \ldots, N$.

Appendix C.1 Proof of Lemma 3

Proof. Denote $v_i(t)=\sum_{j=1}^{N}a_{i,j}(t)\omega_j(t)$ and $\phi_i(t)=P_\Gamma[v_i(t)-\alpha(t)d_i(t)]-(v_i(t)-\alpha(t)d_i(t))$. According to the property of projection, we have

$$||\omega_i(t+1)-\vartheta||^2\leqslant||v_i(t)-\vartheta||^2+\alpha^2(t)||d_i(t)||^2-2\alpha(t)d_i(t)^\top(v_i(t)-\vartheta)-||\phi_i(t)||^2,\quad\forall\vartheta\in\Gamma, \tag{C8}$$

where $\Gamma=\bigcap_{i=1}^{N}\Gamma_i$ is the common domain of ω_i. Since $d_i(t)$ is the subgradient of $\tilde{f}_i$ at $\omega=v_i(t)$, we have

$$d_i(t)^\top(v_i(t)-\vartheta)\geqslant\tilde{f}_i(v_i(t))-\tilde{f}_i(\vartheta). \tag{C9}$$

Combining (C8) and (C9), we obtain

$$\begin{aligned}||\omega_i(t+1)-\vartheta||^2&\leqslant||v_i(t)-\vartheta||^2+\alpha^2(t)||d_i(t)||^2-2\alpha(t)(\tilde{f}_i(v_i(t))-\tilde{f}_i(\vartheta))-||\phi_i(t)||^2\\&\leqslant\sum_{j=1}^{N}a_{i,j}(t)||\omega_j(t)-\vartheta||^2+\alpha^2(t)||d_i(t)||^2-2\alpha(t)(\tilde{f}_i(v_i(t))-\tilde{f}_i(\vartheta))-||\phi_i(t)||^2,\end{aligned} \tag{C10}$$

where the second inequality holds because of the convexity of $||\cdot||^2$. Summing over $i=1,2,\ldots,N$ on both side of (C10), we obtain

$$\sum_{i=1}^{N}||\omega_i(t+1)-\vartheta||^2\leqslant\sum_{j=1}^{N}||\omega_j(t)-\vartheta||^2+\alpha^2(t)\sum_{i=1}^{N}||d_i(t)||^2-2\alpha(t)\sum_{i=1}^{N}(\tilde{f}_i(v_i(t))-\tilde{f}_i(\vartheta))-\sum_{i=1}^{N}||\phi_i(t)||^2. \tag{C11}$$

Appendix C.2 Proof of Lemma 4

Proof. Denote $\bar{\omega}(t)=\frac{1}{N}\sum_{i=1}^{N}\omega_i(t)$, we obtain that

$$\bar{\omega}(t+1)=\bar{\omega}(0)-\frac{1}{N}\sum_{l=0}^{t-1}\sum_{i=1}^{N}\alpha(l)d_i(l)-\frac{\alpha(t)}{N}\sum_{i=1}^{N}d_i(t)+\frac{1}{N}\sum_{l=0}^{t-1}\sum_{j=1}^{N}\phi_j(l)+\frac{1}{N}\sum_{j=1}^{N}\phi_i(t). \tag{C12}$$

With

$$\omega_i(t+1)=P_\Gamma[\sum_{j=1}^{N}a_{i,j}(t)\omega_j(t)-\alpha(t)d_i(t)], \tag{C13}$$

and the transition matrices φ, the following relations can be obtained

$$\begin{aligned}\omega_i(t+1)=&\sum_{j=1}^{N}[\varphi(0:t)]_{i,j}\omega_j(0)-\sum_{l=0}^{t-1}\sum_{j=1}^{N}[\varphi(l+1:t)]_{i,j}\alpha(l)d_j(l)-\alpha(t)d_i(t)+\phi_i(t)\\&+\sum_{l=0}^{t-1}\sum_{j=1}^{N}[\varphi(l+1:t)]_{i,j}\phi_j(l).\end{aligned} \tag{C14}$$

Combining (C12) and (C14), the following relation holds

$$\begin{aligned}||\omega_i(t)-\bar{\omega}(t)||\leqslant&\sum_{l=0}^{t-2}\sum_{j=1}^{N}\left|[\varphi(l+1:t-1)]_{i,j}-\frac{1}{N}\right|||\phi_j(l)||+\sum_{l=0}^{t-2}\sum_{j=1}^{N}\left|[\varphi(l+1:t-1)]_{i,j}-\frac{1}{N}\right|\alpha(l)||d_j(l)||\\&+\alpha(t-1)||d_i(t-1)||+\frac{\alpha(t-1)}{N}\sum_{j=1}^{N}||d_j(t-1)||+\sum_{j=1}^{N}\left|[\varphi(0:t-1)]_{i,j}-\frac{1}{N}\right|||\omega_j(0)||\\&+||\phi_i(t-1)||+\frac{1}{N}\sum_{j=1}^{N}||\phi_j(t-1)||.\end{aligned} \tag{C15}$$

According to Lemma 2, we obtain

$$\begin{aligned}||\omega_i(t)-\bar{\omega}(t)||\leqslant&\sum_{l=0}^{t-2}\sum_{j=1}^{N}\zeta^{-2}\varrho^{t-l-1}||\phi_j(l)||+\sum_{l=0}^{t-2}\sum_{j=1}^{N}\zeta^{-2}\varrho^{t-l-1}\alpha(l)||d_j(l)||+\alpha(t-1)||d_i(t-1)||\\&+\frac{\alpha(t-1)}{N}\sum_{j=1}^{N}||d_j(t-1)||+\zeta^{-2}\varrho^{t}\sum_{j=1}^{N}||\omega_j(0)||+||\phi_i(t-1)||+\frac{1}{N}\sum_{j=1}^{N}||\phi_j(t-1)||.\end{aligned} \tag{C16}$$

Because $||\phi_i(t)||\leqslant L\alpha(t)$, where L is the upper bound of $||\nabla\tilde{f}_i||$ for all $i=1,2,\ldots,N$, (C16) can be changed into

$$||\omega_i(t)-\bar{\omega}(t)||\leqslant C_1\varrho^t+C_2\sum_{l=0}^{t-2}\varrho^{t-l-1}\alpha(l)+4\alpha(t-1)L, \tag{C17}$$

where C_1 and C_2 are both constants, which are defined by $C_1=N\zeta^{-2}\max_j||\omega_j(0)||$ and $C_2=2N\zeta^{-2}L$. Since $0<\varrho<1$ and $\lim_{t\to\infty}\alpha(t)=0$, we obtain $||\omega_i(t)-\bar{\omega}(t)||\to0$, as $t\to\infty$.

Appendix C.3 Proof of Lemma 5

Proof. By multiplying (C16) with $\alpha(t)$ and summing over t, the following relation holds:

$$\begin{aligned}\sum_{t=1}^{\infty}\alpha(t)||\omega_i(t)-\bar{\omega}(t)|| \leqslant & \sum_{t=1}^{\infty}[C_1\alpha(t)\varrho^t + C_2\sum_{l=0}^{t-2}\varrho^{t-l-1}\alpha(t)\alpha(l) + 4\alpha(t)\alpha(t-1)L] \\ \leqslant & \sum_{t=1}^{\infty}[C_1(\varrho^{2t}+\alpha^2(t)) + C_2\sum_{l=0}^{t-2}\varrho^{t-l-1}(\alpha^2(t)+\alpha^2(l)) + 2L(\alpha^2(t)+\alpha^2(t-1))] \\ \leqslant & C_1\sum_{t=1}^{\infty}\varrho^{2t} + \sum_{t=1}^{\infty}(C_3\alpha^2(t)+2L\alpha^2(t-1)) + C_2\sum_{t=1}^{\infty}\sum_{l=0}^{t-2}\varrho^{t-l-1}\alpha^2(l), \end{aligned} \tag{C18}$$

where $C_3 = C_1 + 2L + \frac{C_2\varrho}{1-\varrho}$. Since $0 < \varrho < 1$, and $\sum_{t=1}^{\infty}\alpha^2(t) < \infty$, we obtain $\sum_{t=1}^{\infty}\alpha(t)||\omega_i(t)-\bar{\omega}(t)|| < \infty$.

Appendix D Numerical Simulations

Appendix D.1 Simulations on Synthetic Datasets

In this subsection, all the simulations are on synthetic datasets to illustrate the performance of DGSA. All the input vectors $\{x_i^k\} \in \mathcal{R}^9$ are independently generated from the Gaussian distribution $\mathcal{N}(0,4)$. The true ω is drawn from the Gaussian distribution $\mathcal{N}(0,1)$ and then keep it unchanged. The response value y_i^k is generated from the model $y = \omega^\top x + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,2)$. We assume that there are five agents in the network, and each agent has access to only a portion of dataset $D_i = \{(x_i^k, y_i^k)\}_{k=1}^{m_i}$, where $m_i = 2K$ for all $i = 1, 2, \ldots, 5$. Hence, the number of the whole dataset $\mathcal{D}$ is $10K$.

The goal of all agents in the network is to find the same optimal estimate $\hat{\omega}$ of the true ω. To measure how good the estimate all agents get, the mean square error is defined as follows:

$$MSE = \frac{1}{E}\sum_{e=1}^{E}||\hat{\omega}_e - \omega^*||^2, \tag{D1}$$

where $\hat{\omega}_e$ is the estimation in the e-th experiment, E is the total number of the experiments and ω^* is the optimal solution of the LS problem.

In DGSA, the communication network is assumed to be jointly-connected, which is shown in Figure D1. At time $t = 2n$, the communication topology is assumed to be Figure D1(a), and at time $t = 2n+1$, the topology is assumed to be Figure D1(b), where $n = 1, 2, \ldots$. First, we perform our algorithm just once time and compare the results of DGSA with

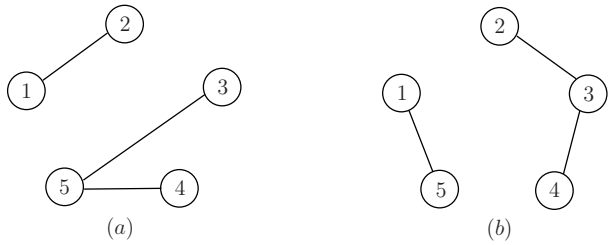

Figure D1 The time-varying topologies of the communication network.

CGSA. To get a pilot estimate ω_0 of ω^*, we choose the parameter $M_{s_0} = 1000$ and apply the uniformly sampling method. The expected sampling size is fixed as $M_s = 2000$ for both algorithms. In DGSA, the common domain of ω in all agents is chosen as $\Gamma = \{\omega : ||\omega - \omega_0|| \leqslant 10\}$. The performance of DGSA and CGSA are shown in Figure D2, which illustrates that all agents obtain a good estimate of ω^*. That means DGSA performs as well as CGSA, even though the datasets are distributed stored in different agents and the communication network is time-varying.

We also investigate the impact of the expected sampling size in DGSA. The expected sampling ratios $r = M_s/M$ are chosen as $0.01, 0.02, \ldots, 0.29, 0.3$. For each sampling ratio, $E = 50$ experiments are simulated for both methods, and the results are illustrated in Figure D3. It is easy to see that as the sampling ratio increases, both methods perform better. When the sampling ratio is chosen as $r = 0.2$, the mean square errors of DGSA and CGSA are nearly equal to 0. We also perform DGSA on different datasets, whose sizes are $M = 10K, 20K$ and $50K$. The expected sampling size and the number of simulations are chosen as $M_s = 0.2M$ and $E = 50$, respectively. The results are listed in Table D1. From Figure D3 and Table D1, we can also see that DGSA has good performance as well as CGSA.

Finally, we compare DGSA with DSA proposed in [4]. The size of dataset we use is $10K$, and the expected sampling size is chosen as $M_s = 2K$. Both of the domains in DGSA and DSA are chosen as $\Gamma = \{\omega : ||\omega|| \leqslant 10\}$. Figure D4 shows the square errors of both algorithms for agent 1 and 2. It is easy to see that both algorithms converge fast and find the optimal solution. We also do $E = 50$ Monte Carlo runs to compare the MSE and average run time. The MSE of DGSA is 2.061e-3, compared with 1.608e-3 of DSA. The run time of DGSA is $0.25s$ per execution including the sampling step compared with $0.40s$ of DSA. The reason why DGSA executes faster and also have a good result is that DGSA only processes a small

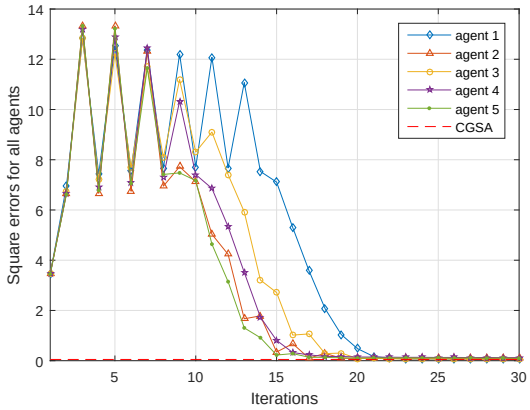

Figure D2 The square errors of DGSA and CGSA.

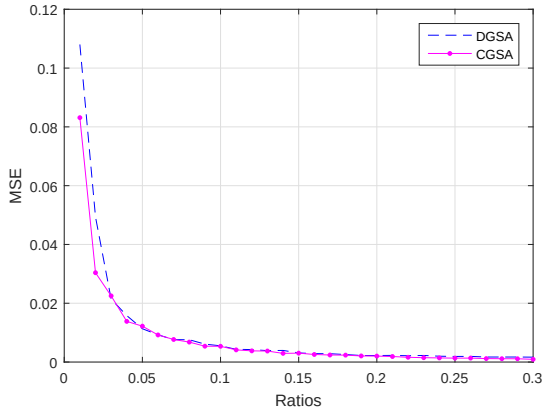

Figure D3 The square errors of DGSA and CGSA.

Table D1 MSEs about 50 Monte Carlo runs on different datasets($r = 0.2$)

Size of Datasets	DGSA	CGSA
10K	2.061e-3	1.552e-3
20K	1.815e-3	9.303e-4
50K	9.486e-4	3.423e-4

Table D2 MSEs about 50 Monte Carlo runs on CASP dataset

Sampling size	45	180	450	1800	4500
DGSA	6.542e-5	4.625e-6	6.573e-6	3.854e-7	1.069e-7
CGSA	2.443e-5	5.314e-6	1.768e-6	3.753e-7	1.132e-7

portion of data points, which are more critical. That is why DGSA can be more suitable to the circumstance that the local datasets in all agents are so large that they cannot deal with by themselves.

Appendix D.2 Simulations on Real Data

In this subsection, we compare the performance of DGSA and CGSA on the real dataset CASP from UCI [5]. The total number of data points in CASP is $M = 45730$ and the dimension of every data is $d = 9$. To get the pilot estimate ω_0 of ω^*, we choose $M_{s_0} = 1K$ data points by uniformly sampling. In DGSA, the domain Γ is determined as $\Gamma = \{\omega : ||\omega - \omega_0|| \leqslant 50\}$, and the communication network is shown as Figure D1. The expected sampling size is chosen as $M_s = 45, 180, 450, 1800, 4500$. The number of simulations is $E = 50$. The MSEs of both methods are shown in Table D2. We can see that DGSA performs as well as CGSA and can adapt to the real circumstances. Also, as the sampling size increases, the results that DGSA and CGSA get are both better.

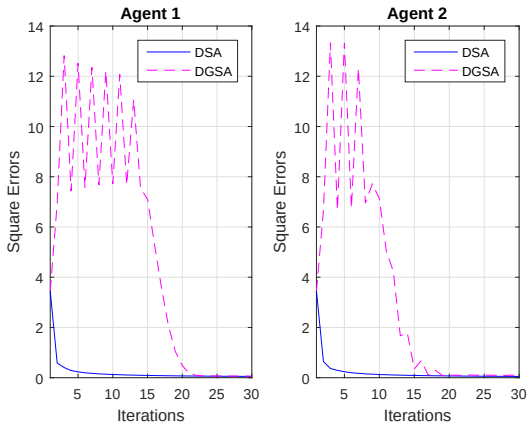

Figure D4 The square errors of DGSA and CGSA.

References

1 Zhu R. Gradient-based sampling: An adaptive importance sampling for least-squares. Advances in Neural Information Processing Systems, 2016: 406-414

2 Tropp J A. User-friendly tools for random matrices: an introduction. CALIFORNIA INST OF TECH PASADENA DIV OF ENGINEERING AND APPLIED SCIENCE, 2012

3 Nedic A and Ozdaglar A. Distributed subgradient methods for multi-agent optimization. IEEE Trans Autom Control, 2009, 54: 48-61

4 Nedic A, Ozdaglar A, Parrilo P A. Constrained consensus and optimization in multi-agent networks. IEEE Trans Automat Contr, 2010, 55: 922-938

5 Dua D, Karra T E. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017