Confidence region for distributed stochastic optimization problem via stochastic gradient tracking method

Shengchao Zhao [a], Yongchao Liu [a]

[a] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*

Abstract

Since stochastic approximation (SA) based algorithms are easy to implement and need less memory, they are very popular in distributed stochastic optimization problems. Many works have focused on the consistency of the objective values and the iterates returned by the SA based algorithms. It is of fundamental interest to know how to quantify the uncertainty associated with SA solutions via the confidence regions of a prescribed level of significance for the true solution. In this paper, we discuss the framework of constructing the asymptotic confidence regions of the optimal solution to distributed stochastic optimization problem with a focus on the distributed stochastic gradient tracking method. To attain this goal, we first present the asymptotic normality of Polyak-Ruppert averaged distributed stochastic gradient tracking method. We then estimate the corresponding covariance matrix through online estimators. Finally, we provide a practical procedure to build the asymptotic confidence regions for the optimal solution. Numerical tests are also conducted to show the efficiency of the proposed methods.

Key words: confidence regions, distributed stochastic optimization, plug-in method, batch-means method, stochastic gradient tracking method

1 Introduction

This paper studies the following distributed stochastic optimization (DSO) problem

$$\min_{x\in\mathbb{R}^d} \; f(x) = \sum_{j=1}^{n} f_j(x), \tag{1}$$

over the undirected networks composed of n agents, where $f_j(x) := \mathbb{E}[g_j(x;\zeta_j)]$ is the local objective function of agent j, ζ_j is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $g_j(x;\zeta_j)$ is a measurable function and $\mathbb{E}[\cdot]$ denotes the expectation with respect to probability $\mathbb{P}$. In problem (1), each agent privately holds the local objective function and can exchange information only with its immediate neighbors. The DSO problem (1) has a wide range of applications, such as large-scale machine learning [6,23,42] and sensor networks [36,3,11], which have been well studied in the past decades [4,40,47].

Algorithms for the distributed optimization problem have been studied extensively in the literature, such as stochastic (sub)gradient descent method [32,5,24,27,43,18,14], dual averaging method [44,15], primal dual method [21,48], stochastic gradient push method [25,2,35]. Recently, many distributed algorithms based on the gradient tracking method have been proposed: [26] introduces a distributed gradient tracking (DGT) method for strong convex optimization and obtains the linear convergence rate of iterates; [31] extends DGT to the distributed stochastic gradient tracking (DSGT) method and shows that the iterates generated by each agent converge to a neighborhood of the optimal solution in a linear rate; [46,33] propose the $\mathcal{AB}$/push-pull method which applies the gradient tracking method to solve a strongly convex optimization problem over a directed graph; [45] discusses the convergence rate of several fundamental algorithmic frameworks, where the stochastic gradient tracking has been combined with variance reduction techniques.

While existing distributed optimization algorithms focus on estimating the optimal solution, less attention has been paid to the statistical inference for the distributed optimization algorithms that update based on random samples. In real-world applications, we are often not just interested in obtaining the optimal solution estimation, but also a measure of the statistical uncertainty associated with the estimation. The statistical inference

Email addresses: zhaoshengchao@mail.dlut.edu.cn (Shengchao Zhao), lyc@dlut.edu.cn (Yongchao Liu).

can provide credibility and validity for the estimation in some critical applications, such as recommender system [1] and autonomous driving [37]. Recently, there are a series of works [10,51,20,8] that study the inferential properties of the stochastic optimization problem equipped with SA based algorithms. Although these works have achieved much progress in this field, all of them focus on the single-machine scenario. In fact, data are usually distributed on different devices in many real-world applications. It is necessary to equip distributed stochastic optimization with inferential capabilities.

The aim of the paper is to investigate the problem of statistical inference of the optimal solution of the DSO problem (1) when SA based algorithm is implemented. We provide computationally efficient methods to build the asymptotic confidence regions of the optimal solutions to the DSO problem (1) when DSGT [31] is employed. Indeed, the confidence regions of the DSO problem have been studied in [10,49]. [10] considers the case that all the agents have the same objective function. As the agents do not need to solve the problem collaboratively, a center may do the statistical inference by collecting all information of agents. [49] uses Sign-Perturbed Sums method to build the non-asymptotic confidence region for the optimal solution to the DSO problem. However, [49] is a region estimation method which is not suitable for the case where the estimation of the optimal solution is needed.

The statistical inference on the optimal solution of DSO problem (1) includes two steps: (I) establish the asymptotic normality for the Polyak-Ruppert averaged [38,29] DSGT and (II) estimate the corresponding covariance matrix in the normal distribution through online estimators. Research on asymptotic normality results for the SA based algorithm can be traced to the works in the 1950s [13,17]. In particular, [29] shows that the averaged SA iterates is asymptotically normal with optimal covariance matrix and optimal convergence rate for strongly convex stochastic optimization problem. For DSO problem without constrains, [5] shows the asymptotic normality of the distributed stochastic gradient descent (DSGD) algorithm under the condition that the weight matrix is doubly stochastic; [24] relaxes the condition on the weight matrix to doubly stochastic in mean sense and establishes the asymptotic normality of the DSGD; [21] focuses on the distributed primal-dual algorithm and presents the asymptotic normality of the solutions. For a DSO with constrains, [39] studies a nonlinear least squares parameter estimation problem and demonstrates the asymptotic normality of solutions when the true parameter x^* is the interior point of the constraint set; [50] studies the stochastic distributed dual average algorithm and shows the asymptotic normality of the iterates when the optimal solution is on the boundary of the constraint set.

To construct an asymptotic confidence region, we need further to construct consistent estimators of covariance matrix in the limit normal distribution. An early work [19] develops a covariance matrix estimator for SA based algorithm by simulating multiple independent replications of iterations. Note that this method needs the historical data, it may lose the advantage of stochastic approximation scheme in terms of data storage. More recently, the seminal work [9] provides two online methods 'plug-in' and 'batch-means' to estimate the covariance matrix when vanilla SGD is implemented on stochastic optimization problems, and shows that the convergence rates of these two methods are close to $\mathcal{O}(\frac{1}{k^{1/2}})$ and $\mathcal{O}\left(\frac{1}{k^{1/8}}\right)$ respectively. [20] extends the batch-means method to the zeroth order stochastic gradient algorithm on stochastic convex optimization problem and obtains the similar convergence rate. [8] employs the plug-in method to do the statistical inference of online decision making and the convergence in probability of the plug-in estimator has been established. To the best of our knowledge, no one has discussed either the plug-in method or the batch-means method for distributed stochastic optimization problems.

As far as we are concerned, the contribution of the paper can be summarized as follows.

1. We establish the asymptotic normality of DSGT. We show that the Polyak-Ruppert averaged DSGT converges to a normality distribution for each agent. The key issue for showing the asymptotic normality of Polyak-Ruppert averaged DSGT is the convergence rate of the iterates of DSGT. By analyzing the intrinsic structure of the accumulated stochastic gradient noise during the step of gradient tracking, we show the stability and agreement of iterates through an extended version of technical result [5, Lemma 3] and then the convergence rate of DSGT. Compared with the convergence rate of DSGT in [31], the new result does not need the boundedness of the variance of local stochastic gradient. Then we present the asymptotic normality of Polyak-Ruppert averaged DSGT by employing the technical tool in [16, Proposition 2], where the asymptotic normality of stochastic dual averaging method has been studied. Different from the asymptotic normality results on DSO problem mentioned above [5,24,21,39,50], the asymptotic normality of Polyak-Ruppert averaged DSGT is based on global stochastic gradient tracker rather than the local stochastic gradient.
2. We present two estimators for the covariance matrix in the limit normal distribution. We extend the plug-in and batch-means methods in [9] to DSGT. For the plug-in method, each agent updates their local estimator by aggregating neighbors' estimator and then plugs the local stochastic gradient product and the second-order derivative in the aggregator at each iteration. The distributed plug-in method does not rely on the gradient tracker, which means it is also suitable for

the distributed stochastic algorithms based on local stochastic gradient. For the batch-means method, although the generalization of the batch means method from SGD to DSGT is straightforward, the proof of its consistency is not straightforward at all, where the tough job is to study the convergence rate of the fourth moment of the iterations and the agreement errors. As far as we know, the extended methods are the first online methods for estimating the covariance matrix for DSO problem.

3. We construct the asymptotic confidence region of the optimal solution of the DSO problem based on the asymptotic normality of DSGT and the two estimators of covariance matrix. Through numerical experiments on ridge regression problem, we conclude that plug-in method needs more information and returns a better estimator of the covariance matrix, batch-means method does not need to communicate with neighbor and returns a relatively rough estimator of covariance matrix.

The rest of this paper is organized as follows. Section 2 presents notation and preliminary conditions on DSO problem. Section 3 studies the convergence rate of iterates and then establishes the asymptotic normality of Polyak-Ruppert averaged DSGT. Section 4 extends plug-in and batch-means methods for vanilla SGD to distributed stochastic gradient tracking method. Numerical results are presented in Section 5 to illustrate the performance of the proposed methods.

2 Notation and preliminary conditions

In this section, we introduce notation and preliminary conditions on stochastic distributed optimizations. $\mathbb{R}^d$ denotes the d-dimension Euclidean space endowed with norm $\|x\| = \sqrt{\langle x, x\rangle}$. Denote $\mathbf{1} := (1\ 1 \dots 1)^\intercal \in \mathbb{R}^n$, $\mathbf{0} = (0\ 0 \dots 0)^\intercal \in \mathbb{R}^d$. $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix. $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of matrix $\mathbf{A}$ and $\mathbf{B}$. For a sequence of random vectors $\{\mu_k\}$ and a random vector μ, $\mu_k \xrightarrow{d} \mu$ denotes the convergence in distribution, $\mathrm{Cov}(\mu)$ denotes the covariance matrix of random vector μ. $N(z, \Sigma)$ is the normal distribution with mean z and covariance matrix Σ. For any sequences $\{a_k\}$ and $\{b_k\}$ of positive numbers, $a_k = \mathcal{O}(b_k)$ if there exists $c > 0$ such that $a_k \le c b_k$ and $a_k \asymp b_k$ if $a_k = \mathcal{O}(b_k)$ and $b_k = \mathcal{O}(a_k)$. For any sequences $\{w_k\}$ and $\{z_k\}$ of random variables, $w_k = \mathcal{O}_p(z_k)$ if for any $\epsilon > 0$ there exists $c > 0$ such that $\mathbb{P}(|w_k/z_k| > c) < \epsilon$ for all $k \ge 0$.

For the distributed optimization problem, the communication relationship between agents is characterized by a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, ..., n\}$ is the node set with node $i \in \mathcal{V}$ representing agent i and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the edge set connecting nodes. $\mathcal{G}$ is an undirected graph if $(i, j) \in \mathcal{E}$ implies that $(j, i) \in \mathcal{E}$. $\mathcal{G}_{\mathbf{A}} = (\mathcal{V}, \mathcal{E}_{\mathbf{A}})$ denotes the graph induced by the nonnegative matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{n\times n}$, where $\mathcal{V} = \{1, 2, ..., n\}$ and $(j, i) \in \mathcal{E}_{\mathbf{A}}$ if and only if $a_{ij} > 0$. Each agent i is able to call a stochastic first-order oracle, which can return a noisy gradient sample of the form $\nabla g_i(x, \zeta_i)$ for the input x.

Next, we recall the distributed stochastic gradient tracking method [31] in Algorithm 1.

Algorithm 1 distributed stochastic gradient tracking method: At each node $i \in \mathcal{V} = \{1, 2, ..., n\}$

Require: initial value $x_{i,0} \in \mathbb{R}^d$, $y_{i,0} = \nabla g_i(x_{i,0}; \zeta_{i,0})$, weight matrices $\mathbf{A} = [a_{ij}]$, stepsize $\alpha_k > 0$.

1: **For** $k = 1, 2, \cdots$ **do**

2: **State update:**

$$x_{i,k+1} = \sum_{j=1}^{n} a_{ij} x_{j,k} - \alpha_k y_{i,k}. \tag{2}$$

3: **Gradient tracking update:**

$$y_{i,k+1} = \sum_{j=1}^{n} a_{ij} y_{j,k} + \nabla g_i(x_{i,k+1}; \zeta_{i,k+1}) - \nabla g_i(x_{i,k}; \zeta_{i,k}),$$

where $\zeta_i, \zeta_{i,0}, \zeta_{i,1}, \cdots$ are independently and identically distributed.

4: **end for**

DSGT is a stochastic gradient variant of distributed gradient tracking method [26]. Different from the distributed stochastic gradient descent method, Step 2 updates the iterates by tracker $y_{i,k}$ rather than $\nabla g_i(x_{i,k}; \zeta_{i,k})$.

Throughout our analysis in the paper, we make the following assumptions. For ease of the explanation of the assumptions, we define the stochastic gradient noise $\epsilon_{i,k} := \nabla g_i(x_{i,k}; \zeta_{i,k}) - \nabla f_i(x_{i,k})$, and the filtration $\mathcal{F}_0 = \sigma\{x_{i,0}, i \in \mathcal{V}\}$,

$$\mathcal{F}_k = \sigma\{x_{i,0}, \epsilon_{i,t} : i \in \mathcal{V}, 0 \le t \le k-1\}, k > 0.$$

Obviously, $x_{i,k}$ and $y_{i,k-1}$ are adapted to $\mathcal{F}_k$.

Assumption 1 (Objective function)(i) $f(x)$ is μ-strongly convex ($\mu > 0$) in x, that is,

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

(ii) $\nabla^2 f(x^*)$ is positive definite and there exists $c > 0$ such that

$$\|\nabla f(x) - \nabla^2 f(x^*)\,(x - x^*)\,\| \le c\|x - x^*\|^2, \quad \forall x \in \mathbb{R}^d, \tag{3}$$

where x^* is the optimal solution to problem (1).

Assumption 2 (Stochastic gradient) For $\forall i \in \mathcal{V}$,

(i) there exists a positive random variable $L_i(\zeta_i)$ such that

$$\|\nabla g_i(x;\zeta_i)-\nabla g_i(y;\zeta_i)\| \leq L_i(\zeta_i)\|x-y\|, \quad \forall x, y \in \mathbb{R}^d;$$

(ii) there exist constants $p \geq 2$ and $c_f > 0$ such that $\mathbb{E}\left[L_i^p(\zeta_i)\right] < \infty$ and

$$\mathbb{E}\left[\|\nabla g_i(x^*;\zeta_i)\|^p\right] \leq c_f^{p/2}. \tag{4}$$

Assumption 3 (Weight matrices and networks)(i) Nonnegative weight matrix $\mathbf{A}$ is doubly stochastic, i.e., $\mathbf{A}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\intercal\mathbf{A} = \mathbf{1}^\intercal$. In addition, $a_{ii} > 0$ for some $i \in \mathcal{V}$.

(ii) The graph $\mathcal{G}_{\mathbf{A}}$ corresponding to the network of agents is undirected and connected.

Remark 1. Assumption 1 (i) guarantees the uniqueness of the optimal solution of DSO problem (1). Assumption 1 (ii) is the standard condition for studying the asymptotic normality of stochastic approximation based algorithms [9,29]. A sufficient condition for the positive definiteness of the second-order derivative is the strong convexity of the objective function, which has been well used to study the convergence rate of DSGT [31,45]. Moreover, (3) holds if $\nabla f(x)$ is globally Lipschitz continuous in x.

Assumption 2 implies the Lipschitz continuity of $\nabla f_i(\cdot)$, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x-y\|,$$

where $L = \max_{1\leq i\leq n} \sqrt[p]{\mathbb{E}[L_i^p(\zeta_i)]}$. From the perspective of the variance of the stochastic gradient, Assumption 2 implies that

$$\mathbb{E}\left[\|\nabla g_i(x;\zeta_i) - \nabla f_i(x)\|^2\right] \quad \leq c_\epsilon\left(1 + \|x-x^*\|^2\right), \tag{5}$$

where c_ϵ is some constant. Obviously, the variance of the stochastic gradient may be unbounded as it is related to x. We need the case $p > 2$ in Assumption 2 (ii) for establishing the asymptotic normality of Polyak-Ruppert averaged DSGT and $p = 4$ for estimating the convergence rate of estimator of covariance matrix.

Assumption 3 implies that $\left(\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal\right)\mathbf{A} = \mathbf{A}\left(\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal\right) = \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$ and the spectral norm ρ of the matrix $\mathbf{A} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$ satisfies $\rho < 1$ [22, Lemma 4].

For ease of catching up with the proof of the main results of paper, we introduce the notations used in the following sections. Denote

$$\begin{aligned}
&x_k := \left[x_{1,k}^\intercal, x_{2,k}^\intercal, \cdots, x_{n,k}^\intercal\right]^\intercal,\\
&y_k := \left[y_{1,k}^\intercal, y_{2,k}^\intercal, \cdots, y_{n,k}^\intercal\right]^\intercal,\\
&\epsilon_k := \left[\epsilon_{1,k}^\intercal, \epsilon_{2,k}^\intercal, \cdots, \epsilon_{n,k}^\intercal\right]^\intercal,\\
&\epsilon_{i,k} := \nabla g_i(x_{i,k};\zeta_{i,k}) - \nabla f_i(x_{i,k}),\\
&\nabla F_k := \left[\nabla f_1\left(x_{1,k}\right)^\intercal, \nabla f_2\left(x_{2,k}\right)^\intercal, \cdots, \nabla f_n\left(x_{n,k}\right)^\intercal\right]^\intercal,\\
&\nabla G_k := \left[\nabla g_1(x_{1,k};\zeta_{1,k})^\intercal, \nabla g_2(x_{2,k};\zeta_{2,k})^\intercal, \cdots,\right.\\
&\qquad\qquad \left.\nabla g_n(x_{n,k};\zeta_{n,k})^\intercal\right]^\intercal,\\
&\bar{x}_k := \left(\frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)x_k, \quad \bar{y}_k := \left(\frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)y_k,
\end{aligned} \tag{6}$$

where x_k, y_k, ∇F_k, ∇G_k and ϵ_k are formed by stacking all agents' iterate, gradient tracker, accurate gradient, stochastic gradient and its noise, $\bar{x}_k$, $\bar{y}_k$ are the average of all agents' iterates and gradient tracker.

Following the notations in (6), Algorithm 1 can be compactly rewritten as

$$x_{k+1} = \tilde{\mathbf{A}}x_k - \alpha_k y_k,\ y_{k+1} = \tilde{\mathbf{A}}y_k + \nabla G_{k+1} - \nabla G_k, \tag{7}$$

where $\tilde{\mathbf{A}} := \mathbf{A}\otimes\mathbf{I}_d$. Throughout this paper, we set stepsize $\alpha_k = a/(k+b)^\alpha$ with $\alpha \in (1/2, 1)$, $a, b > 0$ and $\frac{a}{b^\alpha} \leq \frac{2}{(\mu/n)+L_0}$, where $L_0 = \max_{1\leq i\leq n}\mathbb{E}[L_i(\zeta_i)]$. Moreover, by (7) and the double stochasticity of $\mathbf{A}$,

$$\bar{x}_{k+1} = \bar{x}_k - \alpha_k\bar{y}_k, \tag{8}$$

$$\bar{y}_{k+1} = \frac{1}{n}\sum_{j=1}^n \nabla g_j(x_{j,k+1};\zeta_{j,k+1}). \tag{9}$$

3 The asymptotic normality of DSGT

The asymptotic normality of stochastic approximation based algorithms can be traced to the works of [13,17]. Recently, the asymptotic normality of the distributed stochastic algorithms based on local stochastic gradient have been studied in [5,24,21,39,50]. In this section, we focus on the asymptotic normality of Polyak-Ruppert averaged DSGT. We need to study the agreement and convergence rate of DSGT first. Indeed, the agreement and the convergence rate of DSGT have been discussed in [31]. By setting stepsize $\alpha_k = a/(k+b)$ (a and b are some positive constants) and that the variance of the stochastic gradient is bounded, [31] shows that the decay rates of agreement error and optimality gap in the second-order moment sense are $\mathcal{O}(1/k^2)$ and $\mathcal{O}(1/k)$ respectively. Here, we focus on the case where the variance of the stochastic gradient may be unbounded.

The following lemma is a generalization of [5, Lemma 3] which serves as a technical tool for studying the stabil-

ity and agreement of the distributed stochastic gradient descent algorithm.

Lemma 1 *Suppose that positive sequences $\{\gamma_k\}$, $\{\rho_k\}$, $\{\phi_k\}$ satisfy*

(i) $\{\gamma_k\}$, $\{\rho_k\}$ *are* $[0,1]$*-valued sequences such that* $\sum_{k=0}^{\infty}\gamma_k^2<\infty$, $\limsup_{k\to\infty}\frac{\gamma_k}{\gamma_{k+1}}=1$.

(ii)

$$\begin{aligned}&\limsup_{k\to\infty}\left(\gamma_k\sqrt{\phi_k}+\frac{\phi_{k-1}}{\phi_k}\right)<\infty,\quad \sum_{k=0}^{\infty}\phi_k^{-1}<\infty,\\ &\liminf_{k\to\infty}\left(\gamma_k\sqrt{\phi_k}\right)^{-1}\left(\frac{\phi_{k-1}}{\phi_k}-\rho_k\right)>0.\end{aligned}\tag{10}$$

If sequences $\{u_k\}$, $\{v_k\}$ *satisfy that*

$$\begin{aligned}u_{k+1}\le{}&\rho_k u_k+M\gamma_k\sqrt{u_k(1+u_k+v_k)}\\&+M\gamma_k^2\left(1+\sum_{t=k_0}^{k}\rho^{k-t}u_t+\sum_{t=k_0}^{k}\rho^{k-t}v_t\right),\quad(11)\\ v_{k+1}\le{}&v_k+Mu_k+M\gamma_k\sqrt{u_k(1+u_k+v_k)}\\&+M\gamma_k^2\left(1+\sum_{t=k_0}^{k}\rho^{k-t}u_t+\sum_{t=k_0}^{k}\rho^{k-t}v_t\right),\quad(12)\end{aligned}$$

for $k\ge k_0$*, where scalars* $M>0$, $1>\rho>0$ *and integer* $k_0\ge 1$*, then*

$$\sup_k v_k<\infty,\quad \limsup_{k\to\infty}\phi_k u_k<\infty.$$

Different from the distributed stochastic gradient descent algorithm in [5], DSGT updates the k-th state $x_{i,k}$ based on the tracker $y_{i,k}$. Since $y_{i,k}$ is related to all the past stochastic gradients, the last terms on the right hand side of (11)-(12) are $\sum_{t=k_0}^{k}\rho^{k-t}u_t$ and $\sum_{t=k_0}^{k}\rho^{k-t}v_t$ rather than u_k and v_k in [5, Lemma 3].

The following lemma presents the stability and agreement of DSGT.

Lemma 2 *Suppose that Assumptions 1, 2 (with* $p=2,\ 4$*) and 3 hold. Then there exists positive constant* $\bar{c}$ *such that*

$$\sup_k\mathbb{E}\left[\|\bar{x}_k-x^*\|^p\right]\le\bar{c},\quad \mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^p\right]\le\bar{c}\alpha_k^2.\tag{13}$$

Proof. See Supplementary Materials Section B for the detailed proof. □

Next, we study the convergence rate of the optimal gap $\|\bar{x}_k-x^*\|$.

Theorem 3 (Convergence rate of DSGT) *Suppose that Assumptions 1, 2 (with p=2, 4) and 3 hold. Then there exists constant c such that*

$$\mathbb{E}\left[\|\bar{x}_k-x^*\|^p\right]\le c\alpha_k^{p/2}.$$

Proof. Recall inequality (B.3)

$$\begin{aligned}&4\mathbb{E}\big[\,\|\bar{x}_k-x^*-\alpha_k/n\nabla f(\bar{x}_k)\|^2\times\\&\quad\langle\bar{x}_k-x^*-\alpha_k/n\nabla f(\bar{x}_k),\alpha_k\left(\bar{y}_k-1/n\nabla f(\bar{x}_k)\right)\rangle\,\big]\\&\le 4\mathbb{E}\Bigg[\left\|\bar{x}_k-x^*-\frac{\alpha_k}{n}\nabla f(\bar{x}_k)\right\|^3\times\\&\qquad\left\|\alpha_k\left(\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(x_{j,k})-\frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|\Bigg]\end{aligned}\tag{14}$$

in Supplementary Materials Section B. By (14) and Young's inequality,

$$\begin{aligned}&4\mathbb{E}\Big[\|\bar{x}_k-x^*-\alpha_k/n\nabla f(\bar{x}_k)\|^2\times\\&\quad\langle\bar{x}_k-x^*-\alpha_k/n\nabla f(\bar{x}_k),\alpha_k\left(\bar{y}_k-1/n\nabla f(\bar{x}_k)\right)\rangle\Big]\\&\le 3\tau\mathbb{E}\left[\left\|\bar{x}_k-x^*-\frac{\alpha_k}{n}\nabla f(\bar{x}_k)\right\|^4\right]\\&\quad+\tau^{-3}\mathbb{E}\left[\left\|\alpha_k\left(\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(x_{j,k})-\frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|^4\right].\end{aligned}\tag{15}$$

Substitute (15) into inequality (B.2) in Supplementary Materials Section B,

$$\begin{aligned}&\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^4\right]\\&\le(1+7\tau)\left(1-\frac{\mu}{n}\alpha_k\right)^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\\&\quad+\left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\|\alpha_k\left(\bar{y}_k-1/n\nabla f(\bar{x}_k)\right)\|^4\right]\\&\quad+\tau^{-3}\mathbb{E}\left[\left\|\alpha_k\left(\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(x_{j,k})-\frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|^4\right]\\&\le(1+7\tau)\left(1-\frac{\mu}{n}\alpha_k\right)^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\\&\quad+27\left(3+\frac{4}{\tau}\right)\alpha_k^4\left(\frac{L^4}{n}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\right.\\&\quad\left.+2L^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+c_f^2\right)\\&\quad+\frac{\tau^{-3}\alpha_k^4L^4}{n}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k)\|^4\right]\\&\le(1+7\tau)\left(1-\frac{\mu}{n}\alpha_k\right)^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+\frac{\tau^{-3}\alpha_k^4L^4}{n}\bar{c}\alpha_k^2\\&\quad+27\left(3+\frac{4}{\tau}\right)\alpha_k^4\left(\frac{L^4}{n}\bar{c}\alpha_k^2+2L^4\bar{c}+c_f^2\right),\end{aligned}$$

where c_f and $\bar{c}$ are defined in Assumption 2 and Lemma 2 respectively, the first inequality follows from the fact $\left\|x - x^* - \frac{1}{n}\alpha_k \nabla f(x)\right\| \leq \left(1 - \frac{\mu}{n}\alpha_k\right)\|x - x^*\|$ by [34, Lemma 10], the second inequality follows from inequality (B.5) in Supplementary Materials Section B and the Lipschitz continuity of $\nabla f_i(\cdot)$, the third inequality follows from Lemma 2. Let $\tau = \frac{\mu}{7n}\alpha_k$, we have

$$\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^4\right] \leq \left(1 - \frac{\mu}{n}\alpha_k\right)\mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right] + \mathcal{O}(\alpha_k^3).$$

By [28, Lemma 5, page 46], there exists constant c such that $\mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right] \leq c\alpha_k^2$. The proof is complete. □

Theorem 3 establishes the convergence rate of DSGT with $p = 2, 4$ where the variance of the stochastic gradient may be unbounded, which plays a key role in showing the asymptotic normality of DSGT and estimating the asymptotic covariance matrix [9,51]. When the variance of the stochastic gradient is bounded, the convergence rate of DSGT with $p = 2$ has been established in [33, Theorem 2]. As shown in (5), the variance of the stochastic gradient is related to iterates of DSGT, which induces difficulty in showing the stability and agreement of DSGT. We have to provide a tighter upper bound for the stochastic noise accumulated during the gradient tracking step. Then, we may employ Lemma 1 to show the stability and agreement of DSGT, which plays a key role in establishing the convergence rate of DSGT.

With the convergence rate of iterates of DSGT in Theorem 3, we are ready to present the asymptotic normality of DSGT.

Theorem 4 (Asymptotic normality of DSGT) *Suppose that Assumptions 1, 2 (with p>2) and 3 hold. Then, for any $i \in \mathcal{V}$,*

$$\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}(x_{i,t} - x^*) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right), \tag{16}$$

where $\mathbf{H} := \nabla^2 f(x^*)$ *and* $\mathbf{S} := \mathrm{Cov}\left(\sum_{j=1}^n \nabla g_j(x^*; \zeta_j)\right)$.

Proof. By Lemma 2,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}(\bar{x}_t - x^*) - \frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}(x_{i,t} - x^*)\right\|\right] \\
&\leq \frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}\sqrt{\mathbb{E}\left[\|x_t - \mathbf{1}\otimes\bar{x}_t\|^2\right]} \leq \frac{\sqrt{\bar{c}}}{\sqrt{k}}\sum_{t=0}^{k-1}\alpha_t \to 0.
\end{aligned}$$

Then Slutsky's theorem [12, Theorem 1 in Chapter 8.1] implies (16) if

$$\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}(\bar{x}_t - x^*) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right) \tag{17}$$

holds. In what follows, we show (17) by Lemma C.2 in Supplementary Materials Section C.

Firstly, we rewrite the recursion $\bar{x}_k - x^*$ in the form of (C.2) in Lemma C.2. By (8) and (9),

$$\begin{aligned}
&\bar{x}_{k+1} - x^* \\
&= \left(\mathbf{I}_d - \alpha_k\frac{1}{n}\nabla^2 f(x^*)\right)(\bar{x}_k - x^*) \\
&\quad - \alpha_k\left(\frac{1}{n}\nabla f(\bar{x}_k) - \frac{1}{n}\nabla^2 f(x^*)(\bar{x}_k - x^*)\right) \\
&\quad - \alpha_k\left(\frac{1}{n}\sum_{j=1}^n \nabla g_j(\bar{x}_k; \zeta_{j,k}) - \frac{1}{n}\nabla f(\bar{x}_k)\right) \\
&\quad - \alpha_k\left(\frac{1}{n}\sum_{j=1}^n \nabla g_j(x_{j,k}; \zeta_{j,k}) - \frac{1}{n}\sum_{j=1}^n \nabla g_j(\bar{x}_k; \zeta_{j,k})\right).
\end{aligned} \tag{18}$$

Denote

$$\begin{aligned}
&\Delta_k = \bar{x}_k - x^*, \quad \mathbf{G} = \frac{1}{n}\nabla^2 f(x^*), \quad \gamma_k = \alpha_k, \\
&\mu_k = -\frac{1}{n}\sum_{j=1}^n \nabla g_j(\bar{x}_k; \zeta_{j,k}) + \frac{1}{n}\nabla f(\bar{x}_k)
\end{aligned} \tag{19}$$

and

$$\begin{aligned}
\eta_k = &-\left(\frac{1}{n}\nabla f(\bar{x}_k) - \frac{1}{n}\nabla^2 f(x^*)(\bar{x}_k - x^*)\right) \\
&-\left(\frac{1}{n}\sum_{j=1}^n \nabla g_j(x_{j,k}; \zeta_{j,k}) - \frac{1}{n}\sum_{j=1}^n \nabla g_j(\bar{x}_k; \zeta_{j,k})\right).
\end{aligned} \tag{20}$$

The linear recursion (18) can be rewritten as

$$\Delta_{k+1} = (\mathbf{I}_d - \gamma_k\mathbf{G})\Delta_k + \gamma_k(\eta_k + \mu_k), \tag{21}$$

which is in the form of (C.2) in Lemma C.2.

Next, we verify the conditions (i)-(v) of Lemma C.2. According to the definition α_k and the strong convexity of $f(\cdot)$, conditions (i), (ii) of Lemma C.2 hold obviously. Denote

$$\begin{aligned}
&\mu_k^{(0)} := -\frac{1}{n}\sum_{j=1}^n \nabla g_j(x^*; \zeta_{j,k}), \\
&\mu_k^{(1)} := -\frac{1}{n}\sum_{j=1}^n \nabla g_j(\bar{x}_k; \zeta_{j,k}) + \frac{1}{n}\nabla f(\bar{x}_k) + \frac{1}{n}\sum_{j=1}^n \nabla g_j(x^*; \zeta_{j,k}).
\end{aligned}$$

Then the martingale difference sequence μ_k can be decomposed into $\mu_k = \mu_k^{(0)} + \mu_k^{(1)}$, where $\left\{\mu_k^{(0)}\right\}$ and

$\left\{\mu_k^{(1)}\right\}$ are all martingale difference sequences. Moreover, by Assumption 2, there exists constant c such that

$$\mathbb{E}\left[\left\|\mu_k^{(0)}\right\|^2 |\mathcal{F}_k\right] \leq c, \quad \mathbb{E}\left[\left\|\mu_k^{(1)}\right\|^2 |\mathcal{F}_k\right] \leq c\|\Delta_k\|^2.$$

Recall Lemma C.3 in Supplementary Materials Section C,

$$\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}\mu_t^{(0)} \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{n^2}\mathbf{S}\right).$$

Then condition (iii) of Lemma C.2 holds.

By Assumption 1 (ii) and the Lipscitz continuity of $\nabla g_j(\cdot, \zeta_j)$,

$$\mathbb{E}\left[\|\eta_k\|\right] \leq \frac{1}{n}\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] + \frac{L}{\sqrt{n}}\mathbb{E}\left[\|x_k - \mathbf{1}\otimes\bar{x}_k\|\right] = \mathcal{O}(\alpha_k).$$

Then $\sum_{t=0}^{\infty}\frac{\mathbb{E}[\|\eta_k\|]}{\sqrt{t+1}} < \infty$. Monotone convergence theorem implies $\sum_{t=0}^{\infty}\frac{\|\eta_k\|}{\sqrt{t+1}} < \infty$. Subsequently, Kronecker lemma [7, Chapter 2.7] implies condition (iv) of Lemma C.2.

By Lemma C.4 in Supplementary Materials Section C, $\Delta_k \to 0$ almost surely. Theorem 3 and monotone convergence theorem imply

$$\sum_{t=0}^{\infty}\frac{\|\bar{x}_t - x^*\|^2}{\sqrt{t+1}} < \infty$$

almost surely. Then, Kronecker lemma induces the condition (v) of Lemma C.2.

Summarize above, all the conditions of Lemma C.2 hold. Then

$$\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}(\bar{x}_t - x^*) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right).$$

The proof is complete. □

Theorem 4 shows that the error between the Polyak-Ruppert average $\frac{1}{k}\sum_{t=0}^{k-1}x_{i,t}$ and the optimal solution x^*, normalized by the square root of the iteration counter, converges to a normal random vector in distribution. Indeed, the asymptotic normality of SA based algorithms can be traced back to the 1950s [13,17]. More recently, the asymptotic normality of distributed stochastic optimization algorithms has been well studied, such as stochastic gradient descent [5], dual averaging [50], and primal-dual [21]. As far as we know, Theorem 4 is the first asymptotic normality theorem for the gradient-tracking based algorithm, which paves the way to construct the confidence regions of the optimal solution of the problem (1). Moreover, the asymptotic normality of the constant stepsize DSGT and other (stochastic) gradient tracking algorithms is an interesting topic. We leave it for future research.

4 Online estimation of asymptotic covariance matrix

Statistical inference is a core topic in statistics, and confidence regions have been widely used to quantify the uncertainty in the estimation of model parameters. The asymptotic normality of Polyak-Ruppert averaged DSGT is the first step in building the confidence regions of the optimal solution to the DSO problem. Next, we present estimators of the asymptotic covariance matrix for building the asymptotic confidence regions. Recently, Chen et al. [9] propose two consistent estimators of the asymptotic covariance matrix of Polyak-Ruppert averaged SGD, namely, the plug-in estimator and the batch-means estimator. As for distributed optimization, each agent can access to local information only, we need to extend the plug-in and batch-means methods in [9] to the DSO problem.

4.1 Plug-in method

The idea of the plug-in estimator is to separately estimate $\mathbf{S}$ and $\mathbf{H}$ in (16) by some $\hat{\mathbf{S}}$ and $\hat{\mathbf{H}}$ and then use $\hat{\mathbf{H}}^{-1}\hat{\mathbf{S}}\hat{\mathbf{H}}^{-1}$ as an estimator of $\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}$. For the DSO problem, since each agent is unable to collect global information $\sum_{j=1}^{n}\nabla g_j(x_{j,k};\zeta_{j,k})$, the plug-in method cannot be used directly. An intuitive way is to replace the global information $\sum_{j=1}^{n}\nabla g_j(x_{j,k};\zeta_{j,k})$ with its tracker $y_{i,k}$. Unfortunately, $y_{i,k}$ does not converge to $\frac{1}{n}\sum_{j=1}^{n}\nabla g_j(x_{j,k};\zeta_{j,k})$ [31, Corollary 2]. This motivates us to focus on the case that the covariance matrix $\mathrm{Cov}\left(\sum_{j=1}^{n}\nabla g_j(x^*;\zeta_j)\right)$ has the separable structure, that is,

$$\mathrm{Cov}\left(\sum_{j=1}^{n}\nabla g_j(x^*;\zeta_j)\right) = \sum_{j=1}^{n}\mathrm{Cov}\left(\nabla g_j(x^*;\zeta_j)\right). \quad (22)$$

Condition (22) ensures that agent i may provide an estimation of $\mathrm{Cov}\left(\nabla g_i(x^*;\zeta_i)\right)$ by local information and then shares the estimation among neighbors through communication, which means we may track the covariance matrix. The sufficient condition for (22) is that the ζ_i and ζ_j are uncorrelated for any different agents i and j. Indeed, uncorrelation of random variables of different agents is the standard assumption in distributed stochastic optimization problems, for example, distributed robust estimation problem [36], distributed

maximum likelihood estimation problem [41] and distributed machine learning problem [10], where the samples or observations of different agents are independent. In this subsection, we assume that (22) holds. Then we may provide a distributed variant of plug-in method for DSO problem, which reads as follows.

For $i \in \mathcal{V}$, set $\mathbf{H}_{i,0}$ and $\mathbf{S}_{i,0}$ as zero matrices and update

$$\begin{aligned}&\mathbf{S}_{i,k}\\&= \frac{k-1}{k}\sum_{j=1}^{n} a_{ij}\mathbf{S}_{j,k-1}\\&\quad+\frac{n}{k}\frac{\nabla g_{i,k}\left(\nabla g_{i,k}-\nabla`g_{i,k-1}\right)^{\intercal}+\left(\nabla g_{i,k}-\nabla g_{i,k-1}\right)\nabla g_{i,k}^{\intercal}}{2},\end{aligned} \tag{23}$$

$$\mathbf{H}_{i,k} = \frac{k-1}{k}\sum_{j=1}^{n} a_{ij}\mathbf{H}_{j,k-1} + \frac{n}{k}\nabla^2 g_{i,k}, \tag{24}$$

where $\nabla g_{i,k} := \nabla g_i(x_{i,k};\zeta_{i,k})$, $\nabla^2 g_{i,k} := \nabla^2 g_i(x_{i,k};\zeta_{i,k})$, $\mathbf{S}_{i,k}$ and $\mathbf{H}_{i,k}$ are the estimators of $\mathbf{S}$ and $\mathbf{H}$ respectively.

Note that the estimator $\mathbf{H}_{i,k}$ relies on the Hessian matrix of the loss function, we need the stability of the second order derivative of the objective function.

Assumption 4 [9] For any $i \in \mathcal{V}$, there exists positive scalars L_1 and L_2 such that

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla^2 g_i(x;\zeta_i)-\nabla^2 g_i(x^*;\zeta_i)\right\|\right] \le L_1\|x-x^*\|,\\&\mathbb{E}[\left\|\nabla^2 g_i(x^*;\zeta_i)\right\|^2] \le L_2.\end{aligned}$$

Theorem 5 *Suppose that Assumptions 1, 2 (with $p = 4$), 3 and 4 hold. Then, for any $i \in \mathcal{V}$,*

$$\mathbb{E}\left[\|\mathbf{S}_{i,k}-\mathbf{S}\|\right] = \mathcal{O}\left(\frac{1}{k^{\alpha/2}}\right) \tag{25}$$

and

$$\mathbb{E}\left[\|\mathbf{H}_{i,k}-\mathbf{H}\|\right] = \mathcal{O}\left(\frac{1}{k^{\alpha/2}}\right), \tag{26}$$

where $\alpha \in (1/2, 1)$.

Proof. We just study (25) as the proof of (26) is similar.

For any $i \in \mathcal{V}$, let $\mathbf{C}_{i,0}$ be the zero matrix and

$$\mathbf{C}_{i,k} = \frac{k-1}{k}\sum_{j=1}^{n} a_{ij}\mathbf{C}_{j,k-1}+\frac{1}{k}n\nabla g_{i,k}\left(\nabla g_{i,k}-\nabla g_{i,k-1}\right)^{\intercal}. \tag{27}$$

By the definition of $\mathbf{S}_{i,k}$ in (23), $\mathbf{S}_{i,k} = \frac{\mathbf{C}_{i,k}+\mathbf{C}_{i,k}^{\intercal}}{2}$. Then by the triangle inequality,

$$\|\mathbf{S}_{i,k}-\mathbf{S}\| \le \|\mathbf{C}_{i,k}-\bar{\mathbf{C}}_k\| + \|\bar{\mathbf{C}}_k-\mathbf{\Lambda}_k\| + \|\mathbf{\Lambda}_k-\mathbf{S}\|,$$

where

$$\begin{aligned}&\bar{\mathbf{C}}_k := \sum_{j=1}^{n}\frac{1}{n}\mathbf{C}_{j,k}, \quad \nabla g_{j,t}^* := \nabla g_j(x^*,\zeta_{j,t}),\\&\mathbf{\Lambda}_k := \frac{1}{k}\sum_{t=1}^{k}\sum_{j=1}^{n}\nabla g_{j,t}^*\left(\nabla g_{j,t}^*-\nabla g_{j,t-1}^*\right)^{\intercal}.\end{aligned} \tag{28}$$

We may finish the proof by studying the convergence rate of $\mathbb{E}\left[\left\|\mathbf{C}_{i,k}-\bar{\mathbf{C}}_k\right\|\right]$, $\mathbb{E}\left[\left\|\bar{\mathbf{C}}_k-\mathbf{\Lambda}_k\right\|\right]$ and $\mathbb{E}\left[\|\mathbf{\Lambda}_k-\mathbf{S}\|\right]$ respectively.

Step 1: The convergence rate of $\mathbb{E}\left[\left\|\mathbf{C}_{i,k}-\bar{\mathbf{C}}_k\right\|\right]$. Denote

$$\mathbf{C}_k := \left[\mathbf{C}_{1,k}^{\intercal}, \mathbf{C}_{2,k}^{\intercal}, \cdots, \mathbf{C}_{n,k}^{\intercal}\right]^{\intercal},$$

(27) can be rewritten compactly as

$$\mathbf{C}_k = \frac{k-1}{k}\tilde{\mathbf{A}}\mathbf{C}_{k-1} + \frac{n}{k+1}\mathbf{W}_k, \tag{29}$$

where

$$\begin{aligned}\mathbf{W}_k = n\Big[&(\nabla g_{1,k}-\nabla g_{1,k-1})\nabla g_{1,k}^{\intercal},\ (\nabla g_{2,k}-\nabla g_{2,k-1})\nabla g_{2,k}^{\intercal},\\&\cdots, (\nabla g_{n,k}-\nabla g_{n,k-1})\nabla g_{n,k}^{\intercal}\Big]^{\intercal}.\end{aligned}$$

Note that $\|\mathbf{D}\| \le \|\mathbf{D}\|_F \le \sqrt{d}\|\mathbf{D}\|$ for $\forall \mathbf{D} \in \mathbb{R}^{nd\times d}$, it is sufficient to study the convergence rate of $\mathbb{E}\left[\left\|\mathbf{C}_k-\mathbf{1}\otimes\bar{\mathbf{C}}_k\right\|_F\right]$.

By the recursion (29) of $\mathbf{C}_k$ and the definition of $\bar{\mathbf{C}}_k$,

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{C}_k-\mathbf{1}\otimes\bar{\mathbf{C}}_k\right\|_F\right]\\&=\mathbb{E}\Bigg[\Bigg\|\frac{k-1}{k}\tilde{\mathbf{A}}\mathbf{C}_{k-1}+\frac{1}{k}\mathbf{W}_k\\&\quad-\left(\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right)\left(\frac{k-1}{k}\tilde{\mathbf{A}}\mathbf{C}_{k-1}+\frac{1}{k}\mathbf{W}_k\right)\Bigg\|_F\Bigg]\\&=\mathbb{E}\Bigg[\Bigg\|\frac{k-1}{k}\left(\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right)\left(\mathbf{C}_{k-1}-\mathbf{1}\otimes\bar{\mathbf{C}}_{k-1}\right)\\&\quad+\frac{1}{k}\left(\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right)\mathbf{W}_k\Bigg\|_F\Bigg]\\&\le\frac{k-1}{k}\left\|\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right\|\mathbb{E}\left[\left\|\mathbf{C}_{k-1}-\mathbf{1}\otimes\bar{\mathbf{C}}_{k-1}\right\|_F\right]\\&\quad+\frac{1}{k}\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right\|\mathbb{E}\left[\|\mathbf{W}_k\|_F\right]\\&\le\frac{k-1}{k}\rho\mathbb{E}\left[\left\|\mathbf{C}_{k-1}-\mathbf{1}\otimes\bar{\mathbf{C}}_{k-1}\right\|_F\right]\\&\quad+\frac{1}{k}\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^{\intercal}}{n}\otimes\mathbf{I}_d\right\|\mathbb{E}\left[\|\mathbf{W}_k\|_F\right],\end{aligned} \tag{30}$$

where ρ is the spectral norm of matrix $\mathbf{A} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}$ and the first inequality follows from $\|\mathbf{C}\mathbf{D}\|_F \leq \|\mathbf{C}\|\|\mathbf{D}\|_F$. By the definition of $\mathbf{W}_k$ right after (29),

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{W}_k\|_F\right] &= \mathbb{E}\left[\sqrt{\sum_{j=1}^n \|n\nabla g_{j,k}(\nabla g_{j,k} - \nabla g_{j,k-1})^\intercal\|_F^2}\right] \\ &\leq 2n^2\sqrt{d}\sup_{j\in\mathcal{V},k\geq 0}\mathbb{E}\left[\|\nabla g_{j,k}\|^2\right].\end{aligned} \tag{31}$$

Denote

$$c_g := \sup_{j\in\mathcal{V},k\geq 0}\mathbb{E}\left[\|\nabla g_{j,k}\|^2\right]. \tag{32}$$

By the Lipschitz continuity of $\nabla g_j(\cdot;\zeta_j)$,

$$\begin{aligned}c_g &\leq 2\sup_{j\in\mathcal{V},k\geq 0}\left(\mathbb{E}\left[L^2\|x_{j,k} - x^*\|^2 + \|\nabla g_j(x^*,\zeta_{j,k})\|^2\right]\right) \\ &< \infty,\end{aligned}$$

where the last inequality follows from Lemma 2 and Theorem 3. Substitute c_g into (31), $\mathbb{E}\left[\|\mathbf{W}_k\|_F\right] \leq 2n^2\sqrt{d}c_g$. Therefore, (30) implies the following inequality recursively,

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{C}_k - \mathbf{1}\otimes\bar{\mathbf{C}}_k\right\|_F\right] \\ &\leq \frac{1}{k}\left(\rho^k\mathbb{E}\left[\left\|\mathbf{C}_0 - \mathbf{1}\otimes\bar{\mathbf{C}}_0\right\|_F\right]\right. \\ &\quad \left.+2\sum_{t=1}^k\rho^{k-t}\left\|\mathbf{I}_{n\times d} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|n^2\sqrt{d}c_g\right) \\ &= \frac{2}{k}\sum_{t=1}^k\rho^{k-t}\left\|\mathbf{I}_{n\times d} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|n^2\sqrt{d}c_g \\ &\leq \frac{2}{k}\frac{\left\|\mathbf{I}_{n\times d} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|n^2\sqrt{d}c_g}{1-\rho}.\end{aligned}$$

Note the boundedness of $\frac{2\left\|\mathbf{I}_{n\times d} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|n^2\sqrt{d}c_g}{1-\rho}$, $\mathbb{E}\left[\left\|\mathbf{C}_k - \mathbf{1}\otimes\bar{\mathbf{C}}_k\right\|_F\right] = \mathcal{O}\left(\frac{1}{k}\right)$.

Step 2: The convergence rate of $\mathbb{E}\left[\left\|\bar{\mathbf{C}}_k - \mathbf{\Lambda}_k\right\|\right]$. By the definition of $\bar{\mathbf{C}}_k$ and recursion of (29),

$$\begin{aligned}\bar{\mathbf{C}}_k &= \left(\frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)\left(\frac{k-1}{k}\tilde{\mathbf{A}}\mathbf{C}_{k-1} + \frac{1}{k}\mathbf{W}_k\right) \\ &= \frac{k-1}{k}\bar{\mathbf{C}}_{k-1} + \frac{1}{k}\sum_{j=1}^n\nabla g_{j,k}\left(\nabla g_{j,k} - \nabla g_{j,k-1}\right)^\intercal \\ &= \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\nabla g_{j,t}\left(\nabla g_{j,t} - \nabla g_{j,t-1}\right)^\intercal,\end{aligned}$$

where the second equality follows from the fact $\left(\frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)\tilde{\mathbf{A}} = \left(\frac{\mathbf{1}^\intercal}{n}\mathbf{A}\right)\otimes\mathbf{I}_d = \frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d$. Then

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{\mathbf{C}}_k - \mathbf{\Lambda}_k\right\|\right] \\ &\leq \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\mathbb{E}\left[\left\|\left(\nabla g_{j,t} - \nabla g_{j,t}^*\right)\left(\nabla g_{j,t} - \nabla g_{j,t-1}\right)^\intercal\right.\right. \\ &\quad \left.\left.-\nabla g_{j,t}^*\left(\nabla g_{j,t}^* - \nabla g_{j,t} + \nabla g_{j,t-1} - \nabla g_{j,t-1}^*\right)^\intercal\right\|\right] \\ &\leq \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\sqrt{\mathbb{E}\left[\left\|\nabla g_{j,t} - \nabla g_{j,t}^*\right\|^2\right]\mathbb{E}\left[\|\nabla g_{j,t} - \nabla g_{j,t-1}\|^2\right]} \\ &\quad + \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\left(2\mathbb{E}\left[\|\nabla g_{j,t}^*\|^2\right]\left(\mathbb{E}\left[\left\|\nabla g_{j,t} - \nabla g_{j,t}^*\right\|^2\right]\right.\right. \\ &\quad \left.\left.+\mathbb{E}\left[\left\|\nabla g_{j,t-1} - \nabla g_{j,t-1}^*\right\|^2\right]\right)\right)^{1/2} \\ &\leq \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\sqrt{2c_gL^2\mathbb{E}\left[\|x_{j,t} - x^*\|^2\right]} \\ &\quad + \frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\left(2c_f^2L^2\mathbb{E}\left[\|x_{j,t} - x^*\|^2\right]\right. \\ &\quad \left.+2c_f^2L^2\mathbb{E}\left[\|x_{j,t-1} - x^*\|^2\right]\right)^{1/2},\end{aligned} \tag{33}$$

where c_g and c_f are defined in (32) and (4) respectively, the second inequality follows from Hölder inequality and the third inequality follows from the Lipschitz continuity of $\nabla g_j(\cdot;\zeta_j)$. Then by Lemma 2 and Theorem 3,

$$\mathbb{E}\left[\left\|\bar{\mathbf{C}}_k - \mathbf{\Lambda}_k\right\|\right] = \frac{n}{k}\sum_{t=1}^k\mathcal{O}(\sqrt{\alpha_t}) = \mathcal{O}\left(\frac{1}{k^{\alpha/2}}\right).$$

Step 3: The convergence rate of $\mathbb{E}\left[\|\mathbf{\Lambda}_k - \mathbf{S}\|\right]$. By the definitions of $\mathbf{\Lambda}_k$ and $\mathbf{S}$,

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{\Lambda}_k - \mathbf{S}\|\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^k\sum_{j=1}^n\nabla g_{j,t}^*\left(\nabla g_{j,t}^* - \nabla g_{j,t-1}^*\right)^\intercal - \sum_{j=1}^n\mathbf{S}_j\right\|\right] \\ &\leq \sum_{j=1}^n\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^k\nabla g_{j,t}^*\left(\nabla g_{j,t}^*\right)^\intercal - \mathbb{E}\left[\nabla g_j^*\left(\nabla g_j^*\right)^\intercal\right]\right\|\right] \\ &\quad + \sum_{j=1}^n\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^k\nabla g_{j,t}^*\left(\nabla g_{j,t-1}^*\right)^\intercal - \nabla f_j^*\left(\nabla f_j^*\right)^\intercal\right\|\right],\end{aligned}$$

where $\mathbf{S}_j = \mathrm{Cov}(\nabla g_j(x^*,\zeta_j)), \nabla g_j^* = \nabla g_j(x^*,\zeta_j), \nabla f_j^* = \nabla f_j(x^*)$. It is sufficient to study the convergence rate of

$$\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^k\nabla g_{j,t}^*\left(\nabla g_{j,t}^*\right)^\intercal - \mathbb{E}\left[\nabla g_j^*\left(\nabla g_j^*\right)^\intercal\right]\right\|\right],$$

$$\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^k\nabla g_{j,t}^*\left(\nabla g_{j,t-1}^*\right)^\intercal - \nabla f_j^*\left(\nabla f_j^*\right)^\intercal\right\|\right]$$

respectively. In fact,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t}^{*}\right)^{\intercal}-\mathbb{E}\left[\nabla g_{j}^{*}\left(\nabla g_{j}^{*}\right)^{\intercal}\right]\right\|^{2}\right]\\
&\leq\mathbb{E}\left(\mathbf{tr}\left(\frac{1}{k}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t}^{*}\right)^{\intercal}-\mathbb{E}\left[\nabla g_{j}^{*}\left(\nabla g_{j}^{*}\right)^{\intercal}\right]\right)^{2}\right)\\
&=\mathbf{tr}\left(\mathbb{E}\left(\frac{1}{k}\sum_{t=1}^{k}\left(\nabla g_{j,t}^{*}\left(\nabla g_{j,t}^{*}\right)^{\intercal}-\mathbb{E}\left[\nabla g_{j}^{*}\left(\nabla g_{j}^{*}\right)^{\intercal}\right]\right)\right)^{2}\right)\\
&=\frac{1}{k}\mathbb{E}\mathbf{tr}\left(\left(\nabla g_{j,0}^{*}\left(\nabla g_{j,0}^{*}\right)^{\intercal}\right)^{2}-\left(\mathbb{E}\left[\nabla g_{j}^{*}\left(\nabla g_{j}^{*}\right)^{\intercal}\right]\right)^{2}\right)\\
&=\mathcal{O}\left(\frac{1}{k}\right),
\end{aligned}
$$

where the first inequality follows from the fact that $\|\mathbf{B}\|^2 \leq \mathbf{tr}(\mathbf{B}^2)$ for any symmetric matrix $\mathbf{B}$, the second equality follows from that for $t_1 \neq t_2$,

$$
\mathbb{E}\left[\nabla g_{j,t_1}^{*}\nabla g_{j,t_1}^{*\intercal}\nabla g_{j,t_2}^{*}\nabla g_{j,t_2}^{*\intercal}\right]=\left(\mathbb{E}\left[\nabla g_{j}^{*}\left(\nabla g_{j}^{*}\right)^{\intercal}\right]\right)^{2}.
$$

In addition,

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{k}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t-1}^{*}\right)^{\intercal}-\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\\
&=\frac{1}{k^{2}}\sum_{t_1=1}^{k}\sum_{t_2=1}^{k}\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\nabla g_{j,t_2}^{*}\left(g_{j,t_2-1}^{*}\right)^{\intercal}\right]\\
&\quad-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\\
&=\frac{1}{k^{2}}\sum_{|t_1-t_2|>1}^{k}\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\nabla g_{j,t_2}^{*}\left(g_{j,t_2-1}^{*}\right)^{\intercal}\right]\\
&\quad+\frac{1}{k^{2}}\sum_{|t_1-t_2|=1}^{k}\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\nabla g_{j,t_2}^{*}\left(g_{j,t_2-1}^{*}\right)^{\intercal}\right]\\
&\quad+\mathbb{E}\frac{1}{k^{2}}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t-1}^{*}\right)^{\intercal}\nabla g_{j,t}^{*}\left(\nabla g_{j,t-1}^{*}\right)^{\intercal}\\
&\quad-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\\
&=\frac{2k-2}{k^{2}}\left(\mathbb{E}\left[\nabla g_{j,1}^{*}\left(\nabla g_{j,0}^{*}\right)^{\intercal}\nabla g_{j,2}^{*}\left(\nabla g_{j,1}^{*}\right)^{\intercal}\right]\right.\\
&\quad\left.-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\right)\\
&\quad+\frac{1}{k}\left(\mathbb{E}\left[\nabla g_{j,0}^{*}\left(\nabla g_{j,0}^{*}\right)^{\intercal}\nabla g_{j,0}^{*}\left(g_{j,0}^{*}\right)^{\intercal}\right]\right.\\
&\quad\left.-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\right),
\end{aligned}
$$

where the first equality follows from the fact

$$
\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\right]=\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal},
$$

the third equality follows from the fact

$$
\begin{aligned}
&\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\nabla g_{j,t_2}^{*}\left(\nabla g_{j,t_2-1}^{*}\right)^{\intercal}\right]\\
&=\mathbb{E}\left[\nabla g_{j,t_1}^{*}\left(\nabla g_{j,t_1-1}^{*}\right)^{\intercal}\right]\mathbb{E}\left[\nabla g_{j,t_2}^{*}\left(\nabla g_{j,t_2-1}^{*}\right)^{\intercal}\right]\\
&=\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}
\end{aligned}
$$

for $|t_1 - t_2| > 1$. Then,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{k}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t-1}^{*}\right)^{\intercal}-\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right\|^{2}\right]\\
&\leq\mathbf{tr}\left(\mathbb{E}\left(\frac{1}{k}\sum_{t=1}^{k}\nabla g_{j,t}^{*}\left(\nabla g_{j,t-1}^{*}\right)^{\intercal}-\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\right)\\
&=\frac{2k-2}{k^{2}}\mathbb{E}\mathbf{tr}\left(\nabla g_{j,1}^{*}\left(\nabla g_{j,0}^{*}\right)^{\intercal}\nabla g_{j,2}^{*}\left(\nabla g_{j,1}^{*}\right)^{\intercal}\right.\\
&\quad\left.-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\right)^{2}\\
&\quad+\frac{1}{k}\mathbb{E}\mathbf{tr}\left(\nabla g_{j,0}^{*}\left(\nabla g_{j,0}^{*}\right)^{\intercal}\nabla g_{j,0}^{*}\left(g_{j,0}^{*}\right)^{\intercal}\right.\\
&\quad\left.-\left(\nabla f_{j}^{*}\left(\nabla f_{j}^{*}\right)^{\intercal}\right)^{2}\right)=\mathcal{O}\left(\frac{1}{k}\right),
\end{aligned}
$$

where the last equality follows from the boundedness of $\mathbb{E}\left[\|\nabla g_j(x^*;\zeta_j)\|^4\right]$ for $\forall j \in \mathcal{V}$.

Summarizing the three steps above, (25) holds. The proof is complete. □

Theorem 5 shows that the estimations $\mathbf{H}_{i,k}$ and $\mathbf{S}_{i,k}$ of all agents converge to the same limit $\mathbf{H}$ and $\mathbf{S}$ respectively. Following from Theorem 5, we could obtain the consistency result of the proposed plug-in estimator $\mathbf{H}_{i,k}^{-1}\mathbf{S}_{i,k}\mathbf{H}_{i,k}^{-1}$.

Theorem 6 (Convergence rate of plug-in method) *Suppose that Assumptions 1, 2 (with p=4), 3-4 hold. Then for any $i \in \mathcal{V}$,*

$$
\left\|\mathbf{H}_{i,k}^{-1}\mathbf{S}_{i,k}\mathbf{H}_{i,k}^{-1}-\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right\|=\mathcal{O}_{p}\left(\frac{1}{k^{\alpha/2}}\right),
$$

where $\alpha \in (1/2, 1)$.

Theorem 6 illustrates that the distributed plug-in estimator is consistent, that is, $\mathbf{H}_{i,k}^{-1}\mathbf{S}_{i,k}\mathbf{H}_{i,k}^{-1}$ converges to $\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}$ in probability. Similar to the plug-in estimator for SGD [9, Corollary 4.3], distributed plug-in estimator achieves convergence rate $\mathcal{O}_p\left(\frac{1}{k^{\alpha/2}}\right)$. In order to avoid the possible singularity of $\mathbf{H}_{i,k}$ from statistical randomness, we can replace $\mathbf{H}_{i,k}$ with a nonsingular estimator by using a thresholding scheme introduced in [9], and obtain the convergence rate $\mathcal{O}\left(\frac{1}{k^{\alpha/2}}\right)$ in expectation. Note that the distributed plug-in method does

not rely on the gradient tracker, it is also suitable for the distributed stochastic algorithms which are based on local stochastic gradient.

4.2 Batch-means method

Different from the plug-in method, the batch-means method only uses the iterates from DSGT. Note that Lemma 2 has shown the agreement of the iterates of agents, that is,

$$\mathbb{E}\left[\|x_{i,k}-\bar{x}_k\|^4\right] \longrightarrow 0, \forall i \in \mathcal{V}.$$

Therefore, each agent should be able to estimate the covariance matrix by using individual iterates.

For the fixed agent $i \in \mathcal{V}$, let $\{x_{i,0}, x_{i,1}, \cdots, x_{i,k-1}\}$ be the sequence of iterates generated by DSGT. We split them into $M+1$ batches with the size $m_0, m_1, \cdots, m_M$:

$$\underbrace{\{x_{i,s_0}, \cdots, x_{i,e_0}\}}_{0-\text{th batch}}, \underbrace{\{x_{i,s_1}, \cdots, x_{i,e_1}\}}_{1-\text{th batch}}, \cdots, \underbrace{\{x_{i,s_M}, \cdots, x_{i,e_M}\}}_{M-\text{th batch}},$$

where s_l and e_l are the starting and ending index of $l-$th batch with $s_0 = 0, s_l = e_{l-1}+1$, and $e_M = k-1$, the l-th batch satisfies $\sum_{t=s_l}^{e_l} \alpha_t \asymp N$ and $N = \frac{k^{1-\alpha}}{M+1}$.[1] Then the batch-mean estimator [9] of $\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}$ is given by

$$\frac{1}{M}\sum_{l=1}^{M} m_l \left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)\left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)^{\intercal}, \quad (34)$$

where

$$\hat{X}_{i,m_l} := \frac{1}{m_l}\sum_{t=s_l}^{e_l} x_{i,t}, \quad \hat{X}_{i,M} := \frac{1}{e_M - e_0}\sum_{t=s_1}^{e_M} x_{i,t}, \quad (35)$$

and $m_l = e_l - s_l + 1$.

The generalization of the batch means method from SGD to DSGT is straightforward, but the proof of its consistency is not straightforward at all, which requires the convergence rate of the fourth moment of the iterations and the agreement errors.

Theorem 7 (Convergence rate of batch-means method) *Suppose that Assumptions 1, 2 (with p=4) and 3 hold. Then for any $i \in \mathcal{V}$,*

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M} m_l \left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)\left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)^{\intercal} \right.\right.\\ &\left.\left. -\mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right\|\right]\\ &= \mathcal{O}\left(N^{-\frac{1}{2}} + M^{-\frac{1}{2}} + (NM)^{\frac{-\alpha}{4-4\alpha}}\right),\end{aligned} \quad (36)$$

where $\alpha \in (1/2, 1)$.

Proof. Define two auxiliary sequences $\{\delta_{i,t}\}$ and $\{U_t\}$ as

$$\begin{aligned}&\delta_{i,t} := (x_{i,t} - x^*) - U_t,\\ &U_{t+1} = U_t - \alpha_t \frac{1}{n}\mathbf{H}U_t - \alpha_t \frac{1}{n}\sum_{j=1}^{n}\epsilon_{j,t}, \quad U_0 = \bar{x}_0 - x^*.\end{aligned}$$

By the similar analysis as the proof of [9, Theorem 4.3],

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M} m_l \left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)\left(\hat{X}_{i,m_l} - \hat{X}_{i,M}\right)^{\intercal} - \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right\|\right]\\ &\le \mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M} m_l \left(\hat{U}_{m_l} - \hat{U}_{M}\right)\left(\hat{U}_{m_l} - \hat{U}_{M}\right)^{\intercal} - \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right\|\right]\\ &\quad + \frac{1}{M}\sum_{l=1}^{M} m_l \mathbb{E}\mathbf{tr}\left[\left(\hat{\delta}_{i,m_l} - \hat{\delta}_{i,M}\right)\left(\hat{\delta}_{i,m_l} - \hat{\delta}_{i,M}\right)^{\intercal}\right]\\ &\quad + \frac{2}{M}\sqrt{\sum_{l=1}^{M} m_l \mathbb{E}\mathbf{tr}\left[\left(\hat{U}_{m_l} - \hat{U}_{M}\right)\left(\hat{U}_{m_l} - \hat{U}_{M}\right)^{\intercal}\right]} \times\\ &\quad \sqrt{\sum_{l=1}^{M} m_l \mathbb{E}\mathbf{tr}\left[\left(\hat{\delta}_{i,m_l} - \hat{\delta}_{i,M}\right)\left(\hat{\delta}_{i,m_l} - \hat{\delta}_{i,M}\right)^{\intercal}\right]},\end{aligned} \quad (37)$$

where $\hat{U}_{m_l}$ and $\hat{U}_M$ are defined as in (35) with $x_{i,t}$ being replaced by U_t, $\hat{\delta}_{i,m_l}$ and $\hat{\delta}_{i,M}$ are defined as in (35) with $x_{i,t}$ being replaced by $\delta_{i,t}$. Then, we could finish the proof by studying the convergence rate of the three terms on the right hand side of (37) respectively.

Through mimicking the proof of [9, Lemma 4.7] by replacing martingale difference ξ_k in SGD iterates with the new martingale difference sequence $\bar{\epsilon}_k := \frac{1}{n}\sum_{j=1}^{n}\epsilon_{j,k}$, the first term on the right hand side of (37)

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{M}\sum_{l=1}^{M} m_l \left(\hat{U}_{m_l} - \hat{U}_{M}\right)\left(\hat{U}_{m_l} - \hat{U}_{M}\right)^{\intercal} - \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}\right\|\right]\\ &= \mathcal{O}\left(N^{-\frac{1}{2}} + M^{-\frac{1}{2}} + (NM)^{\frac{-\alpha}{4-4\alpha}}\right).\end{aligned} \quad (38)$$

Note that (38) implies the boundedness of

$$\frac{1}{M}\sum_{l=1}^{M} m_l \mathbb{E}\mathbf{tr}\left[\left(\hat{U}_{m_l} - \hat{U}_{M}\right)\left(\hat{U}_{m_l} - \hat{U}_{M}\right)^{\intercal}\right].$$

By the definitions of $\hat{\delta}_{i,m_l}$ and $\hat{\delta}_{i,M}$,

[1] In this paper, we set $N \asymp M \asymp k^{\frac{1-\alpha}{2}}$.

$$
\begin{aligned}
&\sum_{l=1}^{M} m_l \mathbf{tr}\left(\hat{\delta}_{i,m_l}-\hat{\delta}_{i,M}\right)\left(\hat{\delta}_{i,m_l}-\hat{\delta}_{i,M}\right)^{\intercal} \\
&=\sum_{l=1}^{M} m_l\left(\mathbf{tr}\left(\hat{\delta}_{i,m_l}\hat{\delta}_{i,m_l}^{\intercal}\right)-\mathbf{tr}\left(\hat{\delta}_{i,M}\hat{\delta}_{i,M}^{\intercal}\right)\right) \\
&\leq \sum_{l=1}^{M} m_l \mathbf{tr}\left(\hat{\delta}_{i,m_l}\hat{\delta}_{i,m_l}^{\intercal}\right). \qquad (39)
\end{aligned}
$$

Then, a proper bound for $\frac{1}{M}\sum_{l=1}^{M} m_l \mathbb{E}\left[\left\|\hat{\delta}_{i,m_l}\right\|^2\right]$ will provide the convergence rate of the last two terms on the right hand side of (37).

Denote $\delta_t := (\bar{x}_t - x^*) - U_t$ and $\hat{\delta}_{m_l} := \frac{1}{m_l}\sum_{t=s_l}^{e_l}\delta_t$, we have

$$
\begin{aligned}
&m_l \mathbb{E}\left[\|\hat{\delta}_{i,m_l}\|^2\right] \\
&= m_l \mathbb{E}\left[\left\|\hat{\delta}_{m_l}+\frac{1}{m_l}\sum_{t=s_l}^{e_l}(x_{i,t}-\bar{x}_t)\right\|^2\right] \\
&\leq 2m_l \mathbb{E}\left[\left\|\hat{\delta}_{m_l}\right\|^2\right]+2m_l \mathbb{E}\left[\left\|\frac{1}{m_l}\sum_{t=s_l}^{e_l}(x_{i,t}-\bar{x}_t)\right\|^2\right].
\end{aligned}
$$

On the one hand, by the similar analysis in studying the bound of $m_l \mathbb{E}\left[\left\|\hat{\delta}_{m_l}\right\|^2\right]$ in [9, page 31-32 in Supplement Materal],

$$
m_l \mathbb{E}\left[\|\hat{\delta}_{m_l}\|^2\right] = \mathcal{O}\left(m_l^{-1}s_l^{\alpha}+m_l s_l^{-2\alpha}\right).
$$

On the other hand, by Lemma 2,

$$
\begin{aligned}
m_l \mathbb{E}\left[\left\|\frac{1}{m_l}\sum_{t=s_l}^{e_l}(x_{i,t}-\bar{x}_t)\right\|^2\right] &\leq \sum_{t=s_l}^{e_l}\mathbb{E}\left[\|x_{i,t}-\bar{x}_t\|^2\right] \\
&\leq \sum_{t=s_l}^{e_l}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^2\right] \\
&\leq \bar{c}\sum_{t=s_l}^{e_l}\alpha_t^2 = \mathcal{O}\left(m_l s_l^{-2\alpha}\right).
\end{aligned}
$$

Then

$$
\begin{aligned}
&\frac{1}{M}\sum_{l=1}^{M} m_l \mathbb{E}\left[\|\hat{\delta}_{i,m_l}\|^2\right] \\
&= \mathcal{O}\left(\frac{1}{M}\sum_{l=1}^{M}\left(m_l^{-1}s_l^{\alpha}+m_l s_l^{-2\alpha}\right)\right) \\
&= \mathcal{O}\left(\frac{1}{M}\sum_{l=1}^{M}\left(N^{-1}+l^{-\frac{\alpha}{1-\alpha}}N^{\frac{1-2\alpha}{1-\alpha}}\right)\right) \\
&= \mathcal{O}\left(N^{-1}+M^{-1}\right), \qquad (40)
\end{aligned}
$$

where the second equality follows from the facts $s_l \asymp (lN)^{\frac{1}{1-\alpha}}$ and $m_l \asymp l^{\frac{\alpha}{1-\alpha}}N^{\frac{1}{1-\alpha}}$ by [9, Lemma D1 in Supplement Materal], the third equality follows from the facts $\alpha > 1/2$ and $\frac{1}{M}\sum_{l=1}^{M} l^{-\frac{\alpha}{1-\alpha}} = \mathcal{O}\left(M^{-1}\right)$.

Summarizing (37)-(40), (36) holds. The proof is complete. □

Recall that $N \asymp M \asymp k^{\frac{1-\alpha}{2}}$, the batch-means method achieves the convergence rate of $\mathcal{O}\left(k^{-\frac{1-\alpha}{4}}\right)$, which is slower than that of the plug-in method. On the other hand, the batch-means method has less computation and storage cost as it does not need to calculate the stochastic Hessian matrix. The plug-in estimator is online, which returns immediate estimations of $\mathbf{H}$ and $\mathbf{S}$ at each iteration after receiving the noise observations of the gradient and Hessian matrix. The batch-means method is not a fully online method as the construction of the covariance matrix estimator requires the information on the total number of iterations. It is hard to establish a simple algebraic relation between the successive estimators. More recently, [51] proposes a fully online batch-means method to estimate the asymptotic covariance matrix of the average SGD solution. Extending the fully online batch-means method [51] to the distributed stochastic optimization problem will be the subject of our future works.

5 Experimental results

In this section, we report some preliminary numerical results on the confidence regions of the optimal solution to distributed stochastic optimization problem. According to the asymptotic distribution as given in Theorem 4,

$$
\left\{y : (y-\hat{x})^{\intercal}\Sigma^{-1}(y-\hat{x}) \leq \frac{1}{k}\chi_{\beta}^2(d)\right\}
$$

defines an approximate $1-\beta$ confidence region for the optimal solution to DSO problem (1), where $\Sigma = \mathbf{H}^{-1}\mathbf{S}\mathbf{H}^{-1}$, $\hat{x} := \frac{1}{k}\sum_{t=0}^{k-1}x_{i,t}$, $\chi_{\beta}^2(d)$ is defined as the number that satisfies $P\left(U > \chi_{\beta}^2(d)\right) = \beta$ for a χ^2 random variable U with d degrees of freedom.

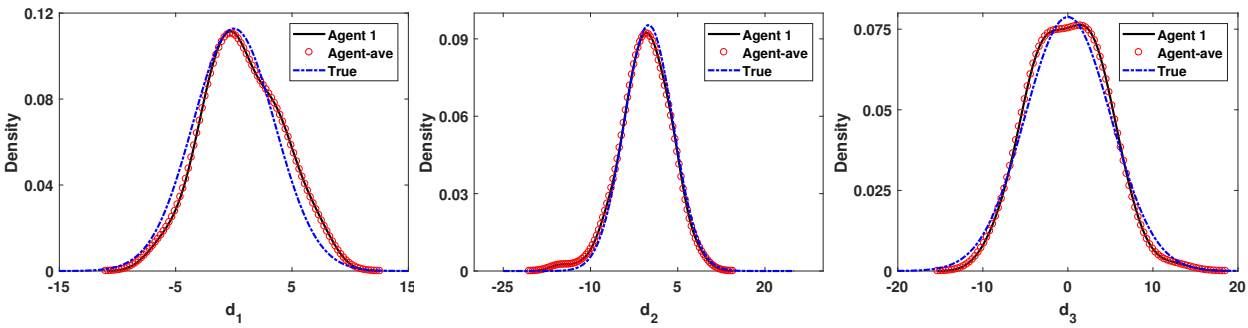

Fig. 1. Density of components.

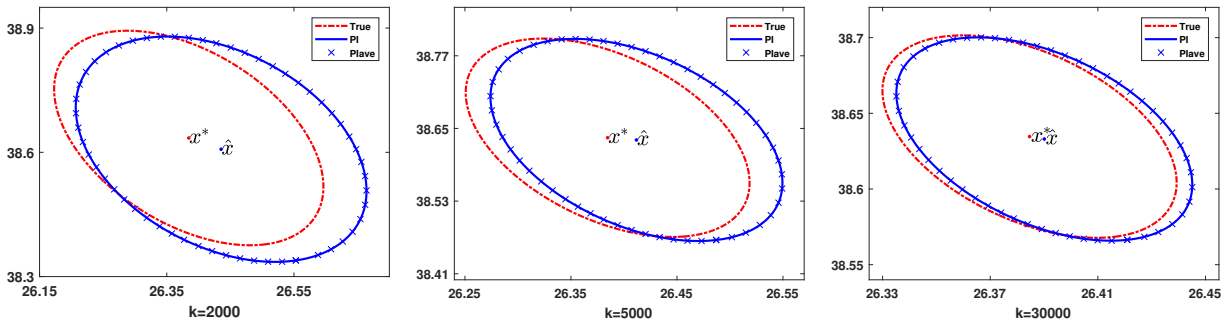

Fig. 2. Cross section of approximate confidence region (PI).

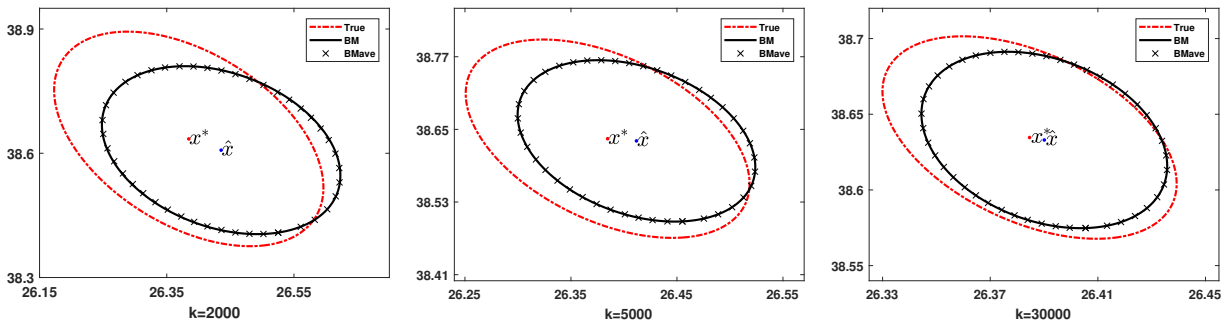

Fig. 3. Cross section of approximate confidence region (BM).

We report the empirical performance of the proposed methods through the ridge regression problem [31]:

$$\min_{x\in\mathbb{R}^d} \; f(x) = \sum_{j=1}^{n} \left(\mathbb{E}\left[\left(w_j^\intercal x - v_j \right)^2 \right] + \gamma \|x\|^2 \right), \qquad (41)$$

where $f_i(x) := \mathbb{E}\left[\left(w_i^\intercal x - v_i \right)^2 \right] + \gamma\|x\|^2$ is the objective function of agent i. In problem (41), each agent $\in \mathcal{V}$ has access to sample (w_i, v_i) given by the following linear model

$$v_i = w_i^\intercal \tilde{x}_i + \nu_i,$$

where w_i is the regression vector accessible to agent i, ν_i is the observation noise of agent i and $\tilde{x}_i$ is unknown parameter. Assume that random variables w_i and ν_i are independent and then the unique solution is $x^* = \left(\sum_{j=1}^n \mathbb{E}[w_j w_j^\intercal] + n\gamma \mathbf{I} \right)^{-1} \sum_{j=1}^n \mathbb{E}[w_j w_j^\intercal] \tilde{x}_j$.

In this experiment, $d = 3, n = 20$, for all $i \in \mathcal{V}$, random variables $w_i \in [0, 20]^3$ are uniformly distributed, ν_i are drawn from the Gaussian distribution $N(0, 1)$, parameters $\tilde{x}_i = [1,\ 50,\ 100]^\intercal$ and $\gamma = 1$. For the weight matrices, we use the similar setting of network topology in [30]. An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is generated by adding random links to a ring network, where a link exists between any two nonadjacent nodes with a probability $p > 0.7$. The weights a_{ji} are defined by the

Metropolis rule [40]:

$$a_{ji} = \begin{cases} \dfrac{1}{\max\{n_i, n_j\} + 1} & j \in \mathcal{N}_i, \\ 1 - \sum\limits_{j \in \mathcal{N}_i} \mathbf{A}_{ji} & j = i, \\ 0 & \text{otherwise}, \end{cases} \tag{42}$$

where $\mathcal{N}_i := \{j | (i,j) \in \mathcal{E}, j \neq i\}$ is agent i's set of neighbors, $n_i := |\mathcal{N}_i|$ is the cardinality of $\mathcal{N}_i$. Moreover, the stepsize is $\alpha_k = 0.1/(k+1)^{0.505}$.

We first carry out tests on the asymptotic normality of iterates (Theorem 4). We do 100 Monte-Carlo simulations of running DSGT with 30000 iterations. In Figure 1, the black solid curve and red circle curve denote the estimated density of agent 1 and average of all agents respectively, the blue dash-dot curve denotes the true density. Figure 1 seems to confirm Theorem 4 since we can see that the estimated density of a component of normalized estimation error is close to the density of the limiting normal distribution, which is also confirmed by a Kolmogorov–Smirnov test.

Next, we construct the asymptotic confidence regions of the optimal solution. We record the confidence regions with the nominal coverage probability $1 - \beta$ to 95% and the covariance matrix are constructed by plug-in (PI) and batch-means (BM) methods after $k = 2000$, 5000 and 30000 iterations respectively. For the stability, we perform 50 Monte-Carlo simulations and report the results with the average covariance matrix and the average of iterates. Figure 2 depicts the confidence regions with PI, where the red solid, blue solid and blue cross ellipses denote the confidence regions for agent 1, average of all agents and the true one respectively. As we can observe from Figure 2, the ellipses for agent 1 and the average of all agents are coincide for iteration $k = 2000$, which indicates the agreement of PI. For iteration $k = 30000$, the three confidence regions are almost coincide, which confirms the convergence of PI in Theorem 6. Similar to PI, the agreement of BM is observable for $k = 2000$ iteration in Figure 3. We can conclude that there is a tendency for the convergence of the confidence regions based on BM to true one, which confirms Theorem 7. Obviously, the convergence of PI in Figure 2 is faster than the convergence of BM in Figure 3, which verifies the fact that PI and BM have convergence rates $\mathcal{O}_p(\frac{1}{k^{\alpha/2}})$ and $\mathcal{O}\left(\frac{1}{k^{(1-\alpha)/4}}\right)$ (in expectation) respectively. Just as we commented at the beginning of subsection 4.2, the underlying reason is that BM estimates covariance matrix based on individual iterates rather than communicating additional quantities with their neighbors.

We report the coverage rate of the confidence regions. We perform 500 Monte-Carlo simulations and record the percentage of the 95% confidence regions containing the true solution in Table 1, where the columns of PI, PIave, BM and BMave record the results for agent 1 and the average of all agents by PI and BM respectively. From the Table 1, we can see that PI gives better coverage rate than BM, where the coverage rates at iterations $k = 5000$ and $k = 30000$ are nearly 95%. On the other hand, the coverage rate of BM has a tendency of converging to 95%. Moreover, we can also observe the agreement of PI and BM from the second and the third columns, the fourth and the fifth columns respectively.

Table 1. Coverage rate(%).

Iterations \ Methods	PI	PIave	BM	BMave
2000	89.4	89.6	63	63.4
5000	93.8	94	70	70
30000	94.6	94.8	82	82

Acknowledgements

The research is supported by the NSFC #11971090 and Fundamental Research Funds for the Central Universities DUT22LAB301.

References

[1] Gediminas Adomavicius and Jingjing Zhang. Stability of recommendation algorithms. *ACM Transactions on Information Systems*, 30(4):1–31, 2012.

[2] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 344–353. PMLR, 2019.

[3] Sergio Barbarossa, Stefania Sardellitti, and Paolo Di Lorenzo. Distributed detection and estimation in wireless sensor networks. *arXiv preprint arXiv:1307.1448*, 2013.

[4] Dimitri Bertsekas and John Tsitsiklis. *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., Englewood Cliffs, 1989.

[5] Pascal Bianchi, Gersende Fort, and Walid Hachem. Performance of a distributed stochastic approximation algorithm. *IEEE Transactions on Information Theory*, 59(11):7405–7418, 2013.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[7] Han. Fu Chen. *Stochastic approximation and its applications*, volume 64. Kluwer Academic Publishers, New York, 2006.

[8] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association*, 116(534):708–719, 2021.

[9] Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 02 2020.

[10] Xi Chen, Weidong Liu, and Yichen Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arXiv:1811.11368*, 2021.

[11] Symeon Chouvardas, Konstantinos Slavakis, and Sergios Theodoridis. Adaptive robust distributed learning in diffusion sensor networks. *IEEE Transactions on Signal Processing*, 59(10):4692–4707, 2011.

[12] Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer, New York, 1978.

[13] Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25:463–483, 1954.

[14] Soham De, Gavin Taylor, and Tom Goldstein. Variance reduction for distributed stochastic gradient descent. *arXiv preprint arXiv:1512.01708*, 2017.

[15] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[16] John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21 – 48, 2021.

[17] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[18] Alireza Fallah, Mert Gurbuzbalaban, Asuman Ozdaglar, Umut Simsekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *arXiv preprint arXiv:1910.08701*, 2019.

[19] Ming hua Hsieh and P.W. Glynn. Confidence regions for stochastic approximation algorithms. In *Proceedings of the Winter Simulation Conference*, volume 1, pages 370–376, 2002.

[20] Yanhao Jin, Tesi Xiao, and Krishnakumar Balasubramanian. Statistical inference for polyak-ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint arXiv:2102.05198*, 2021.

[21] Jinlong Lei, Han Fu Chen, and Hai Tao Fang. Asymptotic properties of primal-dual algorithm for distributed stochastic optimization over random networks with imperfect communications. *SIAM Journal on Control and Optimization*, 56(3):2159–2188, 2018.

[22] Xiuxian Li, Lihua Xie, and Yiguang Hong. Distributed aggregative optimization over multi-agent networks. *arXiv preprint arXiv:2005.13436*, 2020.

[23] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3043–3052. PMLR, 2018.

[24] Gemma Morral, Pascal Bianchi, Gersende Fort, and Jérémie Jakubowicz. Distributed stochastic approximation: The price of non-double stochasticity. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers*, pages 1473–1477, 2012.

[25] Angelia Nedić and Alex Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

[26] Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[27] Naeimeh Omidvar, Mohammad Ali Maddah-Ali, and Hamed Mahdavi. A hybrid-order distributed sgd method for non-convex optimization to balance communication overhead, computational complexity, and convergence rate. *arXiv preprint arXiv:2003.12423*, 2020.

[28] Boris T. Polyak. *Introduction to Optimization*. Optimization Software, NY, 1987.

[29] Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[30] Shi Pu. A robust gradient tracking method for distributed optimization over directed networks. In *2020 59th IEEE Conference on Decision and Control*, pages 2335–2341, 2020.

[31] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

[32] Shi Pu, Alex Olshevsky, and Ioannis Ch. Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *arXiv preprint arXiv:1906.02702*, 2021.

[33] Shi Pu, Wei Shi, Jinming Xu, and Angelia Nedić. A push-pull gradient method for distributed optimization in networks. In *2018 IEEE Conference on Decision and Control*, pages 3385–3390, 2018.

[34] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[35] Muhammad I. Qureshi, Ran Xin, Soummya Kar, and Usman A. Khan. Push-saga: A decentralized stochastic algorithm with variance reduction over directed graphs. *arXiv preprint arXiv:2008.06082*, 2020.

[36] M. Rabbat and R. Nowak. Distributed optimization in sensor networks. In *Third International Symposium on Information Processing in Sensor Networks*, pages 20–27, 2004.

[37] Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 0(0):1–14, 2022.

[38] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[39] Anit Kumar Sahu, Soummya Kar, José M. F. Moura, and H. Vincent Poor. Distributed constrained recursive nonlinear least-squares estimation: Algorithms and asymptotics. *IEEE Transactions on Signal and Information Processing over Networks*, 2(4):426–441, 2016.

[40] Ali H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

[41] Ioannis D. Schizas, Alejandro Ribeiro, and Georgios B. Giannakis. Consensus in ad hoc wsns with noisy links—part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350–364, 2008.

[42] Ohad Shamir and Nathan Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing*, pages 850–857, 2014.

[43] Brian Swenson, Ryan Murray, Soummya Kar, and H. Vincent Poor. Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima. *arXiv preprint arXiv:2003.02818*, 2020.

[44] Konstantinos I. Tsianos and Michael G. Rabbat. Distributed dual averaging for convex optimization under communication delays. In *2012 American Control Conference*, pages 1067–1072, 2012.

[45] Ran Xin, Soummya Kar, and Usman A. Khan. Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence. *IEEE Signal Processing Magazine*, 37(3):102–113, 2020.

[46] Ran Xin and Usman A. Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

[47] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H. Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[48] Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H. Johansson. A primal-dual sgd algorithm for distributed nonconvex optimization. *arXiv preprint arXiv:2006.03474*, 2020.

[49] Vincenzo Zambianchi, Michel Kieffer, Francesca Bassi, Gianni Pasolini, and Davide Dardari. Distributed sps algorithms for non-asymptotic confidence region evaluation. In *2014 European Conference on Networks and Communications*, pages 1–5, 2014.

[50] Shengchao Zhao, Xing Min Chen, and Yongchao Liu. Asymptotic properties of dual averaging algorithm for constrained distributed stochastic optimization. *arXiv preprint arXiv:2009.02740*, 2020.

[51] Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association*, 118(541):393–404, 2023.

Supplementary Materials for Confidence Region for Distributed Stochastic Optimization Problem via Stochastic Gradient Tracking Method

Shengchao Zhao and Yongchao Liu*

A Some technical results of proof of Lemma 2

Lemma A.1. *Suppose that Assumptions 1, 2 (with p=2, 4) and 3 hold. Then there exists a constant c_e such that*

$$\mathbb{E}[\|\xi_k\|^p] \leq c_e^{p/2}\left(1+\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^p\right]+\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^p\right]\right). \tag{A.1}$$

Proof. We just study the case $p=4$ as the analysis for case $p=2$ is similar.

Recall that

$$\begin{aligned}\xi_k &= \tilde{\mathbf{A}}\xi_{k-1}+\epsilon_k-\epsilon_{k-1}\\ &= \sum_{t=0}^{k-1}\tilde{\mathbf{A}}^{k-1-t}(\tilde{\mathbf{A}}-\mathbf{I}_{n\times d})\epsilon_t+\epsilon_k\\ &= \sum_{t=0}^{k}\tilde{\mathbf{A}}(k,t)\epsilon_t,\end{aligned}$$

*School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China, e-mail: zhaoshengchao@mail.dlut.edu.cn (Shengchao Zhao), lyc@dlut.edu.cn (Yongchao Liu)

where $\tilde{\mathbf{A}}(k,t) := \tilde{\mathbf{A}}^{k-1-t}(\tilde{\mathbf{A}} - \mathbf{I}_{n\times d})$ for $t < k$ and $\tilde{\mathbf{A}}(k,k) = \mathbf{I}_{n\times d}$. Then

$$\begin{aligned}
\mathbb{E}[\|\xi_k\|^4] &= \mathbb{E}\left[(\langle \xi_k, \xi_k\rangle)^2\right] \\
&= \sum_{t_1=0}^{k}\sum_{t_2=0}^{k}\sum_{t_3=0}^{k}\sum_{t_4=0}^{k} \mathbb{E}\left[\nu(k,t_1)^\intercal \nu(k,t_2)\nu(k,t_3)^\intercal \nu(k,t_4)\right] \\
&\le \sum_{t_1=0}^{k}\sum_{t_2=0}^{k}\sum_{t_3=0}^{k}\sum_{t_4=0}^{k} \mathbb{E}\left[\|\nu(k,t_1)\|\|\nu(k,t_2)\|\|\nu(k,t_3)\|\|\nu(k,t_4)\|\right] \\
&\le \sum_{t_1=0}^{k}\sum_{t_2=0}^{k}\sum_{t_3=0}^{k}\sum_{t_4=0}^{k} \Big(\left\|\tilde{\mathbf{A}}(k,t_1)\right\| \left\|\tilde{\mathbf{A}}(k,t_2)\right\| \left\|\tilde{\mathbf{A}}(k,t_3)\right\| \times \\
&\qquad \left\|\tilde{\mathbf{A}}(k,t_4)\right\| \mathbb{E}\left[\|\epsilon_{t_1}\|\|\epsilon_{t_2}\|\|\epsilon_{t_3}\|\|\epsilon_{t_4}\|\right]\Big),
\end{aligned}$$

where $\nu(k,t) = \tilde{\mathbf{A}}(k,t)\epsilon_t$. By Assumption 3,

$$\begin{aligned}
\left\|\mathbf{A}^k(\mathbf{A} - \mathbf{I}_n)\right\| &= \left\|\left(\mathbf{A} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\right)\mathbf{A}^{k-1}(\mathbf{A} - \mathbf{I}_n)\right\| \\
&\le \rho\left\|\mathbf{A}^{k-1}(\mathbf{A} - \mathbf{I}_n)\right\| \le \rho^k \|\mathbf{A} - \mathbf{I}_n\|.
\end{aligned}$$

Denoting $c_a = \max\left\{1, \frac{\|\mathbf{A}-\mathbf{I}_n\|}{\rho}\right\}$,

$$\left\|\tilde{\mathbf{A}}(k,t)\right\| \le c_a \rho^{k-t}. \tag{A.2}$$

Subsequently,

$$\begin{aligned}
\mathbb{E}[\|\xi_k\|^4] &\le c_a^4 \sum_{t_1=0}^{k}\sum_{t_2=0}^{k}\sum_{t_3=0}^{k}\sum_{t_4=0}^{k} \rho^{4k-t_1-t_2-t_3-t_4}\mathbb{E}\left[\|\epsilon_{t_1}\|\|\epsilon_{t_2}\|\|\epsilon_{t_3}\|\|\epsilon_{t_4}\|\right] \\
&\le c_a^4 \sum_{t_1=0}^{k}\sum_{t_2=0}^{k}\sum_{t_3=0}^{k}\sum_{t_4=0}^{k} \left(\rho^{4k-t_1-t_2-t_3-t_4}\frac{\mathbb{E}\left[\|\epsilon_{t_1}\|^4 + \|\epsilon_{t_2}\|^4 + \|\epsilon_{t_3}\|^4 + \|\epsilon_{t_4}^4\|\right]}{4}\right) \\
&\le \frac{c_a^4}{(1-\rho)^3}\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\epsilon_t\|^4\right].
\end{aligned} \tag{A.3}$$

Recall the definition of $\epsilon_t := \left[\epsilon_{1,t}^\intercal, \epsilon_{2,t}^\intercal, \cdots, \epsilon_{n,t}^\intercal\right]^\intercal$,

$$\begin{aligned}
\mathbb{E}\left[\|\epsilon_t\|^4\right] &\le n\sum_{j=1}^{n}\mathbb{E}\left[\|\epsilon_{j,t}\|^4\right] \\
&= n\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla g_j(x_{j,t};\zeta_{j,t}) - \nabla g_j(x^*;\zeta_{j,t}) + \nabla f_j(x^*)\right. \\
&\qquad \left. -\nabla f_j(x_{j,t}) + (\nabla g_j(x^*;\zeta_{j,t}) - \nabla f_j(x^*))\|^4\right] \\
&\le 54nL^4\mathbb{E}\left[\|x_t - \mathbf{1}\otimes x^*\|^4\|\right] + 216n^2c_f^2 \\
&\le 432nL^4\mathbb{E}\left[\|x_t - \mathbf{1}\otimes \bar{x}_t\|^4\|\right] + 432n^2L^4\mathbb{E}\left[\|\bar{x}_t - x^*\|^4\|\right] + 216n^2c_f^2,
\end{aligned} \tag{A.4}$$

where c_f is defined in Assumption 2 (ii), the second inequality follows from Assumption 2. Substitute (A.4) into (A.3),

$$\begin{aligned}
&\mathbb{E}[\|\xi_k\|^4] \\
&\leq \frac{c_a^4}{(1-\rho)^4}\sum_{t=0}^{k}\rho^{k-t}\left(432nL^4\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\|\right]\right. \\
&\quad \left.+432n^2L^4\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\|\right]+216n^2c_f^2\right) \\
&\leq c_e^2\left(1+\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]+\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right]\right),
\end{aligned}$$

where $c_e=\frac{c_a^4}{(1-\rho)^4}\max\left\{232n^2L^4,\frac{216n^2c_f^2}{1-\rho}\right\}$. The proof is complete. □

B Proof of Lemma 2

Proof. We just study the case $p=4$ as the case $p=2$ is similar. We employ Lemma 1 to prove

$$\sup_k \mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\leq\bar{c},\quad \mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\leq\bar{c}\alpha_k^2. \tag{B.1}$$

and the proof can be finished by the following two steps: find relationships of u_k and v_k in the forms of (10) and (11) in Lemma 1 firstly, and then verify conditions (i)-(ii) of Lemma 1.

Step 1. Let θ_1 and θ_2 be any random vectors and τ be any positive scalar. Obviously,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_1+\theta_2\|^4\right] &= \mathbb{E}\left[\left(\|\theta_1\|^2+\|\theta_2\|^2+2\langle\theta_1,\theta_2\rangle\right)^2\right] \\
&= \mathbb{E}\left[\|\theta_1\|^4+\|\theta_2\|^4+2\|\theta_1\|^2\|\theta_2\|^2+4\|\theta_1\|^2\langle\theta_1,\theta_2\rangle\right. \\
&\quad \left.+4\|\theta_2\|^2\langle\theta_1,\theta_2\rangle+4\left(\langle\theta_1,\theta_2\rangle\right)^2\right] \\
&\leq \mathbb{E}\left[\|\theta_1\|^4+\|\theta_2\|^4+6\|\theta_1\|^2\|\theta_2\|^2+4\|\theta_1\|^2\langle\theta_1,\theta_2\rangle\right. \\
&\quad \left.+4\|\theta_2\|^2(\|\theta_1\|\|\theta_2\|)\right] \\
&\leq \mathbb{E}\left[\|\theta_1\|^4+3\|\theta_2\|^4+8\|\theta_1\|^2\|\theta_2\|^2+4\|\theta_1\|^2\langle\theta_1,\theta_2\rangle\right] \\
&\leq (1+4\tau)\mathbb{E}\left[\|\theta_1\|^4\right]+\left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\|\theta_2\|^4\right]+4\mathbb{E}\left[\|\theta_1\|^2\langle\theta_1,\theta_2\rangle\right],
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality $\langle x,y\rangle\leq\|x\|\|y\|$, the second inequality follows from Young's inequality $ab\leq\frac{a^p}{p}+\frac{b^q}{q},\frac{1}{p}+\frac{1}{q}=1$ $(p=2)$, the last inequality follows from the fact $\|\theta_1\|^2\|\theta_2\|^2=\left(\tau^{-0.5}\|\theta_1\|\right)^2\left(\tau^{0.5}\|\theta_2\|\right)^2$ and Young's inequality $(p=2)$. Choosing

$$\theta_1=\bar{x}_k-x^*-\alpha_k1/n\nabla f(\bar{x}_k),\quad \theta_2=-\alpha_k(\bar{y}_k-1/n\nabla f(\bar{x}_k)),$$

we have $\bar{x}_{k+1} - x^* = \theta_1 + \theta_2$ and

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^4\right] \\
&\leq (1+4\tau)\mathbb{E}\left[\|\bar{x}_k - x^* - \alpha_k/n\nabla f(\bar{x}_k)\|^4\right] + \left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\|\alpha_k\left(\bar{y}_k - 1/n\nabla f(\bar{x}_k)\right)\|^4\right] \\
&\quad + 4\mathbb{E}\big[\,\|\bar{x}_k - x^* - \alpha_k/n\nabla f(\bar{x}_k)\|^2 \left\langle \bar{x}_k - x^* - \alpha_k/n\nabla f(\bar{x}_k), \alpha_k\left(\bar{y}_k - 1/n\nabla f(\bar{x}_k)\right)\right\rangle\big]. \qquad \text{(B.2)}
\end{aligned}
$$

For the third term on the right hand side of (B.2),

$$
\begin{aligned}
&4\mathbb{E}\big[\,\|\bar{x}_k - x^* - \alpha_k/n\nabla f(\bar{x}_k)\|^2 \times \\
&\quad \left\langle \bar{x}_k - x^* - \alpha_k/n\nabla f(\bar{x}_k), \alpha_k\left(\bar{y}_k - 1/n\nabla f(\bar{x}_k)\right)\right\rangle\big] \\
&= 4\mathbb{E}\left[\left\|\bar{x}_k - x^* - \frac{\alpha_k}{n}\nabla f(\bar{x}_k)\right\|^2 \left\langle \bar{x}_k - x^* - \frac{\alpha_k}{n}\nabla f(\bar{x}_k), \alpha_k\left(\frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k}) - \frac{1}{n}\nabla f(\bar{x}_k)\right)\right\rangle\right] \\
&\leq 4\mathbb{E}\left[\left\|\bar{x}_k - x^* - \frac{\alpha_k}{n}\nabla f(\bar{x}_k)\right\|^3 \left\|\alpha_k\left(\frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k}) - \frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|\right], \qquad \text{(B.3)}
\end{aligned}
$$

where the equality follows from the fact

$$
\mathbb{E}\left[\bar{y}_k \middle| \mathcal{F}_k\right] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \nabla g(x_{j,k};\zeta_{j,k}) \middle| \mathcal{F}_k\right] = \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k}).
$$

Substitute (B.3) into (B.2),

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^4\right] \\
&\leq (1+4\tau)\left(1-\frac{\alpha_k\mu}{n}\right)^4 \mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right] + \left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\left\|\alpha_k\left(\bar{y}_k - \frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|^4\right] \\
&\quad + 4\left(1-\frac{\alpha_k\mu}{n}\right)^3 \mathbb{E}\left[\|\bar{x}_k - x^*\|^3 \left\|\alpha_k\left(\frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k}) - \frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|\right] \\
&\leq \left[(1+4\tau)\left(1-\frac{\alpha_k\mu}{n}\right)^4 + 4\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3\right]\mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right] + \left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\|\alpha_k\left(\bar{y}_k - 1/n\nabla f(\bar{x}_k)\right)\|^4\right] \\
&\quad + 4/\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3 \frac{\alpha_k^2 L^2}{n}\mathbb{E}\left[\|\bar{x}_k - x^*\|^2 \|x_k - \mathbf{1}\otimes\bar{x}_k\|^2\right] \\
&\leq \left[(1+4\tau)\left(1-\frac{\alpha_k\mu}{n}\right)^4 + 4\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3\right]\mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right] + \left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\left\|\alpha_k\left(\bar{y}_k - \frac{1}{n}\nabla f(\bar{x}_k)\right)\right\|^4\right] \\
&\quad + \frac{\alpha_k^2 L^2}{n}\frac{4}{\tau}\left(1-\frac{\alpha_k\mu}{n}\right)^3 \sqrt{\mathbb{E}\left[\|\bar{x}_k - x^*\|^4\right]\mathbb{E}\left[\|x_k - \mathbf{1}\otimes\bar{x}_k\|^4\right]}, \qquad \text{(B.4)}
\end{aligned}
$$

where the first inequality follows from the fact $\left\|x - x^* - \frac{1}{n}\alpha_k\nabla f(x)\right\| \leq \left(1-\frac{\mu}{n}\alpha_k\right)\|x - x^*\|$ by [4, Lemma 10], the second inequality follows from the fact $a^3 b \leq \tau a^4 + \frac{a^2 b^2}{\tau}$ and the Lipschitz continuity of $\nabla f_i(\cdot)$, the third inequality follows from the Hölder inequality. For the second term

on the right hand side of (B.4),

$$
\begin{aligned}
&\left(3+\frac{4}{\tau}\right)\mathbb{E}\left[\|\alpha_k\left(\bar{y}_k-1/n\nabla f(\bar{x}_k)\right)\|^4\right]\\
&=\left(3+\frac{4}{\tau}\right)\alpha_k^4\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^n\left(\nabla g_j(x_{j,k};\zeta_{j,k})-\nabla g_j(x^*;\zeta_{j,k})\right)\right.\right.\\
&\quad\left.\left.-\frac{1}{n}\sum_{j=1}^n\left(\nabla f_j(\bar{x}_k)-\nabla f_j(x^*)\right)+\frac{1}{n}\sum_{j=1}^n\nabla g_j(x^*;\zeta_{j,k})\right\|^4\right]\\
&\le 27\left(3+\frac{4}{\tau}\right)\alpha_k^4\frac{1}{n}\sum_{j=1}^n\left(\mathbb{E}\left[\|\nabla g_j(x_{j,k};\zeta_{j,k})-\nabla g_j(x^*;\zeta_{j,k})\|^4\right]\right.\\
&\quad\left.+\mathbb{E}\left[\|\nabla f_j(\bar{x}_k)-\nabla f_j(x^*)\|^4\right]+\mathbb{E}\left[\|\nabla g_j(x^*;\zeta_{j,k})\|^4\right]\right)\\
&\le 27\left(3+\frac{4}{\tau}\right)\alpha_k^4\left(\frac{L^4}{n}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]+L^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+c_f^2\right),
\end{aligned}
\tag{B.5}
$$

where c_f is defined in Assumption 2, the equality follows from the fact $\frac{1}{n}\sum_{j=1}^n\nabla f_j(x^*)=0$, the second inequality follows from the Lipschitz continuity of $\nabla f_i(\cdot)$ and $\nabla g_i(\cdot;\zeta_i)$. Substitute (B.5) into (B.4),

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^4\right]\\
&\le\left[(1+4\tau)\left(1-\frac{\alpha_k\mu}{n}\right)^4+4\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3\right]\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\\
&\quad+27\left(3+\frac{4}{\tau}\right)\alpha_k^4\left(\frac{L^4}{n}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\right.\\
&\quad\left.+L^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+c_f^2\right)+4/\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3\frac{\alpha_k^2L^2}{n}\sqrt{\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]}.
\end{aligned}
\tag{B.6}
$$

Let $\tau=\frac{\alpha_k\mu}{8n}$, then for any $k\ge 0$,

$$
\left[(1+4\tau)\left(1-\frac{\alpha_k\mu}{n}\right)^4+4\tau\left(1-\frac{\alpha_k\mu}{n}\right)^3\right]=\left(1-\frac{\alpha_k^2\mu^2}{2n^2}\right)\left(1-\frac{\alpha_k\mu}{n}\right)^3<\left(1-\frac{\alpha_k\mu}{n}\right)^3.
$$

Subsequently,

$$
\begin{aligned}
&\mathbb{E}\left[\|\bar{x}_{k+1}-x^*\|^4\right]\\
&\le\left(1-\frac{\alpha_k\mu}{n}\right)^3\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+\frac{32L^2}{\mu}\alpha_k\sqrt{\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]}\\
&\quad+27\left(3\alpha_0+\frac{32n}{\mu}\right)\alpha_k^3\left(\frac{L^4}{n}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]+L^4\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+c_f^2\right).
\end{aligned}
\tag{B.7}
$$

On the other hand, for any vectors $\theta_1,\ \theta_2$ and positive scalar τ,

$$\begin{aligned}
\|\theta_1+\theta_2\|^4 &\le \left(\|\theta_1\|^2+\|\theta_2\|^2+2\|\theta_1\|\|\theta_2\|\right)^2\\
&= \|\theta_1\|^4+\|\theta_2\|^4+4\|\theta_1\|^3\|\theta_2\|+4\|\theta_2\|^3\|\theta_1\|+6\|\theta_1\|^2\|\theta_2\|^2\\
&= \|\theta_1\|^4+\|\theta_2\|^4+4\left(\left(\frac{\tau}{3}\right)^{\frac{1}{4}}\|\theta_1\|\right)^3\left(\left(\frac{\tau}{3}\right)^{-\frac{3}{4}}\|\theta_2\|\right)\\
&\quad+4\left(\tau^{\frac{1}{4}}\|\theta_1\|\right)\left(\tau^{-\frac{1}{12}}\|\theta_2\|\right)^3+6\left(\left(\frac{\tau}{3}\right)^{\frac{1}{4}}\|\theta_1\|\right)^2\left(\left(\frac{\tau}{3}\right)^{-\frac{1}{4}}\|\theta_2\|\right)^2 \qquad \text{(B.8)}\\
&\le (1+3\tau)\|\theta_1\|^4+\left(1+\frac{27}{\tau^3}+\frac{3}{\tau^{1/3}}+\frac{9}{\tau}\right)\|\theta_2\|^4, \qquad \text{(B.9)}
\end{aligned}$$

where the second inequality follows from Young's inequality ($p=4/3, 4/3$ and 2 for the last three terms in (B.8) respectively). Choosing

$$\begin{aligned}
\theta_1 &= \left(\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)(x_k-\mathbf{1}\otimes\bar{x}_k),\\
\theta_2 &= -\alpha_k\left(\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)y_k
\end{aligned}$$

in (B.9), by the equality $x_{k+1}=\tilde{\mathbf{A}}x_k-\alpha_k y_k$ and the double stochasticity of $\mathbf{A}$, we have $x_{k+1}-\mathbf{1}\otimes\bar{x}_{k+1}=\theta_1+\theta_2$ and

$$\begin{aligned}
\|x_{k+1}-\mathbf{1}\otimes\bar{x}_{k+1}\|^4 &\le (1+3\tau)\left\|\left(\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)(x_k-\mathbf{1}\otimes\bar{x}_k)\right\|^4\\
&\quad+\left(1+\frac{27}{\tau^3}+\frac{3}{\tau^{1/3}}+\frac{9}{\tau}\right)\alpha_k^4\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|^4\|y_k\|^4. \qquad \text{(B.10)}
\end{aligned}$$

Denote auxiliary sequences $\{\xi_k\}$ and $\{y_k^{'}\}$ as

$$\begin{aligned}
\xi_k &= \tilde{\mathbf{A}}\xi_{k-1}+\epsilon_k-\epsilon_{k-1}, \quad \xi_0=\epsilon_0,\\
y_{k+1}^{'} &= \tilde{\mathbf{A}}y_k^{'}+\nabla F_{k+1}-\nabla F_k, \quad y_0^{'}=\nabla F_0.
\end{aligned}$$

By the definitions of y_k and $y_k^{'}$, we have $y_k=y_k^{'}+\xi_k$ and that

$$\begin{aligned}
\mathbb{E}\left[\|y_k\|^4\right] &= \mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}-\mathbf{1}\otimes\left(1/n\nabla f(\bar{x}_k)-\bar{y}_k^{'}\right)+\mathbf{1}\otimes 1/n\nabla f(\bar{x}_k)+\xi_k\right\|^4\right]\\
&\le 64\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4+\frac{L^4}{n^2}\|\bar{x}_k-x^*\|^4+\|\xi_k\|^4\right]\\
&\le 64\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4+\frac{L^4}{n^2}\|\bar{x}_k-x^*\|^4\right]\\
&\quad+64c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]+64c_e^2+64c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right], \qquad \text{(B.11)}
\end{aligned}$$

where $\bar{y}_k^{'}:=\left(\frac{\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)y_k^{'}$, the first inequality follows from the Lipscitz continuity of $\nabla f_j(\cdot)$, the fact $\bar{y}_k^{'}=1/n\sum_{j=1}^n\nabla f_j(x_{j,k})$ and the cr-inequality $\mathbb{E}\left[\left|\sum_{j=1}^n X_j\right|^r\right]\le n^{r-1}\sum_{j=1}^n\mathbb{E}\left[|X_j|^r\right]$ [3,

9.1.a in Chapter 9], the second inequality follows from Lemma A.1. Take expectation on both sides of (B.10) and substitute (B.11) into it,

$$
\begin{aligned}
&\mathbb{E}\left[\|x_{k+1}-\mathbf{1}\otimes\bar{x}_{k+1}\|^4\right]\\
&\leq \frac{1+\rho^4}{2}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]+c_1\alpha_k^4\left(\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\right.\\
&\quad+\frac{L^4}{n^2}\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right]+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]+c_e^2\\
&\quad\left.+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right]\right),
\end{aligned}
\tag{B.12}
$$

where $\rho<1$ is the spectral norm of $\mathbf{A}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$, the inequality follows from settings $\tau=\frac{1-\rho^4}{6\rho^4}$ and

$$
c_1=64\left(1+\frac{27}{\tau^3}+\frac{3}{\tau^{1/3}}+\frac{9}{\tau}\right)\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|^4.
$$

Choosing

$$
\begin{aligned}
\theta_1&=\left(\tilde{\mathbf{A}}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)\left(y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right),\\
\theta_2&=\left(\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right)\left(\nabla F_{k+1}-\nabla F_k\right)
\end{aligned}
$$

in (B.9), by the definitions of $y_{k+1}^{'}$ and $\bar{y}_{k+1}^{'}$, we have $y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}=\theta_1+\theta_2$ and

$$
\begin{aligned}
\left\|y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}\right\|^4\leq(1+3\tau)\rho^4\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+\left(1+\frac{27}{\tau^3}+\frac{3}{\tau^{1/3}}+\frac{9}{\tau}\right)\times\\
\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|^4\|\nabla F_{k+1}-\nabla F_k\|^4,
\end{aligned}
\tag{B.13}
$$

where ρ is the spectral norm of $\mathbf{A}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$. By the Lipschitz continuity of $\nabla f(\cdot)$,

$$
\begin{aligned}
\|\nabla F_{k+1}-\nabla F_k\|^4&\leq L^4\|x_{k+1}-x_k\|^4\\
&=L^4\left\|\tilde{\mathbf{A}}x_k-\alpha_k y_k-x_k\right\|^4\\
&=L^4\left\|\left(\tilde{\mathbf{A}}-\mathbf{I}_{n\times d}\right)(x_k-\mathbf{1}\otimes\bar{x}_k)-\alpha_k y_k\right\|^4\\
&\leq 8L^4\left\|\tilde{\mathbf{A}}-\mathbf{I}_{n\times d}\right\|^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4+8L^4\alpha_k^4\|y_k\|^4.
\end{aligned}
\tag{B.14}
$$

Substitute (B.14) into (B.13) and set $\tau=\frac{1-\rho^4}{6\rho^4}$,

$$
\begin{aligned}
&\left\|y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}\right\|^4\\
&\leq\frac{1+\rho^4}{2}\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+c_2\left(\left\|\tilde{\mathbf{A}}-\mathbf{I}_{n\times d}\right\|^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4+\alpha_k^4\|y_k\|^4\right),
\end{aligned}
$$

where $c_2 = 8L^4\left(1+\frac{27}{\tau^3}+\frac{3}{\tau^{1/3}}+\frac{9}{\tau}\right)\left\|\mathbf{I}_{n\times d}-\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\otimes\mathbf{I}_d\right\|^4$. Take expectation on both sides of above inequality,

$$\begin{aligned}\mathbb{E}\left[\left\|y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}\right\|^4\right] \leq \frac{1+\rho^4}{2}\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4\right]+c_2\left\|\tilde{\mathbf{A}}-\mathbf{I}_{n\times d}\right\|^4\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\\ +64c_2\alpha_k^4\Bigg(\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4+\frac{L^4}{n^2}\|\bar{x}_k-x^*\|^4\right]\\ +c_e^2+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right]\Bigg), \quad \text{(B.15)}\end{aligned}$$

where the inequality follows from (B.11).

Multiplying $c^{'}=\frac{1-\rho^4}{4c_2\left\|\tilde{\mathbf{A}}-\mathbf{I}_{n\times d}\right\|^4}$ on both sides of inequality (B.15), we have

$$\begin{aligned}c^{'}\mathbb{E}\left[\left\|y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}\right\|^4\right] \leq \frac{1+\rho^4}{2}c^{'}\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4\right]+\frac{1-\rho^4}{4}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\\ +64c^{'}c_2\alpha_k^4\Bigg(\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4+L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right.\\ \left.+\frac{L^4}{n^2}\|\bar{x}_k-x^*\|^2\right]+c_e^2+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]\\ +c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right]\Bigg).\end{aligned}$$

Adding above inequality and (B.12), we have

$$\begin{aligned}&\mathbb{E}\left[\|x_{k+1}-\mathbf{1}\otimes\bar{x}_{k+1}\|^4\right]+c^{'}\mathbb{E}\left[\left\|y_{k+1}^{'}-\mathbf{1}\otimes\bar{y}_{k+1}^{'}\right\|^4\right]\\ &\leq\frac{3+\rho^4}{4}\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]\\ &+\left(\frac{1+\rho^4}{2}+(c_1/c^{'}+64c_2)\alpha_k^4\right)c^{'}\mathbb{E}\left[\left\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\right\|^4\right]\\ &+(c_1+64c_2c^{'})\alpha_k^4\Bigg(\mathbb{E}\left[L^4\|x_k-\mathbf{1}\otimes\bar{x}_k\|^2\right]\\ &+\frac{L^4}{n^2}\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|x_t-\mathbf{1}\otimes\bar{x}_t\|^4\right]\\ &+c_e^2+c_e^2\sum_{t=0}^{k}\rho^{k-t}\mathbb{E}\left[\|\bar{x}_t-x^*\|^4\right]\Bigg). \quad \text{(B.16)}\end{aligned}$$

Denote

$$\begin{aligned}u_k&=\mathbb{E}\left[\|x_k-\mathbf{1}\otimes\bar{x}_k\|^4\right]+c^{'}\mathbb{E}\left[\|y_k^{'}-\mathbf{1}\otimes\bar{y}_k^{'}\|^4\right],\\ v_k&=\mathbb{E}\left[\|\bar{x}_k-x^*\|^4\right], \quad \gamma_k=\alpha_k\end{aligned}$$

and $\rho_k = \max\left\{\frac{3+\rho^4}{4},\ \frac{1+\rho^4}{2} + (c_1/c^{'} + 64c_2)\alpha_k^4\right\}$. Then by inequalities (B.7) and (B.16),

$$u_{k+1} \leq \rho_k u_k + M\gamma_k\sqrt{u_k(1+u_k+v_k)} + M\gamma_k^2\left(1 + \sum_{t=0}^{k}\rho^{k-t}u_t + \sum_{t=0}^{k}\rho^{k-t}v_t\right), \tag{B.17}$$

$$v_{k+1} \leq v_k + M\gamma_k\sqrt{u_k(1+u_k+v_k)} + M\gamma_k^2\left(1 + \sum_{t=0}^{k}\rho^{k-t}u_t + \sum_{t=0}^{k}\rho^{k-t}v_t\right), \tag{B.18}$$

where

$$M = \max\left\{2(c_1 + 64c_2c^{'})\left(L^4 + c_e^2\right), 27\left(3\alpha_0 + \frac{32n}{\mu}\right)(c_f^2 + L^4)\right\}.$$

(B.17) and (B.18) are in the forms of (10) and (11) in Lemma 1.

Step 2. By the definitions of γ_k and ρ_k, there exists positive integer k_0 such that γ_k, ρ_k are $[0,1]$-valued when $k \geq k_0$ (without loss of generality, suppose $k_0 = 0$). Then condition (i) of Lemma 1 holds.

Let $\phi_k = 1/\alpha_k^2$. Obviously,

$$\limsup_{k\to\infty}\left(\gamma_k\sqrt{\phi_k} + \frac{\phi_{k-1}}{\phi_k}\right) = 2, \quad \sum_{k=0}^{\infty}\phi_k^{-1} < \infty,$$
$$\liminf_{k\to\infty}\left(\gamma_k\sqrt{\phi_k}\right)^{-1}\left(\frac{\phi_{k-1}}{\phi_k} - \rho_k\right) = \frac{1-\rho^2}{4} > 0,$$

which implies the condition (ii) of Lemma 1. Then by Lemma 1, $\sup_k \mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^4\right] < \infty$ and

$$\sup_k \frac{1}{\alpha_k^2}\left(\mathbb{E}\left[\|x_k - \mathbf{1}\otimes\bar{x}_k\|^4\right] + c^{'}\mathbb{E}\left[\|y_k^{'} - \mathbf{1}\otimes\bar{y}_k^{'}\|^4\right]\right) < \infty.$$

The proof is complete. □

C Some technical results of proof of Theorem 4

Lemma C.1. *[5] Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\{\mathcal{F}_k\}$ be a nondecreasing sequence of σ-algebra. Let $\{v_k\}$, $\{a_k\}$, $\{b_k\}$ and $\{\phi_k\}$ be sequences of nonnegative random variables adapted to $\mathcal{F}_k$. If $\sum_{k=1}^{\infty} a_k < \infty$, $\sum_{k=1}^{\infty} b_k < \infty$ almost surely and that for all k,*

$$\mathbb{E}[v_{k+1}|\mathcal{F}_k] \leq (1+a_k)v_k + b_k - \phi_k. \tag{C.1}$$

Then $\{v_k\}$ converges to a finite random variable v_∞ and $\sum_{k=1}^{\infty}\phi_k < \infty$ almost surely.

Lemma C.2. *[2, Proposition 2] Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $\{\mathcal{F}_k\}$ be a nondecreasing sequence of σ-algebra. Let $\{\Delta_k\}$, $\{\mu_k\}$ and $\{\eta_k\}$ be sequences of random vectors in $\mathbb{R}^d$, $\{\gamma_k\}$ be the nonnegative scalars. $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ be the identity matrix. If recursion*

$$\Delta_{k+1} = (\mathbf{I}_d - \gamma_k\mathbf{G})\,\Delta_k + \gamma_k\,(\mu_k + \eta_k) \tag{C.2}$$

satisfies following conditions:

(i) $\gamma_k = \mathcal{O}\left(\frac{1}{k^\alpha}\right)$, $\alpha \in (1/2, 1)$.

(ii) $\mathbf{G}$ *is a positive definite matrix.*

(iii) μ_k *has the decomposable structure* $\mu_k = \mu_k^{(0)} + \mu_k^{(1)}$, *where* $\mu_k^{(0)}$ *and* $\mu_k^{(1)}$ *are both martingale sequences adapted to* $\mathcal{F}_{k+1}$. *In addition,*

$$\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \mu_k^{(0)} \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where $\mathbf{0} = (0\ 0 \ldots 0)^\intercal \in \mathbb{R}^d$, *and there exists a constant* c *such that*

$$\mathbb{E}\left[\left\|\mu_k^{(0)}\right\|^2 \middle| \mathcal{F}_k\right] \leq c, \quad \mathbb{E}\left[\left\|\mu_k^{(1)}\right\|^2 \middle| \mathcal{F}_k\right] \leq c\|\Delta_k\|^2. \tag{C.3}$$

(iv) η_k *is adapted to* $\mathcal{F}_{k+1}$ *and* $\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \|\eta_t\| \longrightarrow 0$ *almost surely.*

(v) Δ_k *is adapted to* $\mathcal{F}_k$, $\Delta_k \longrightarrow 0$ *and* $\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \|\Delta_t\|^2 \longrightarrow 0$ *almost surely.*

Then

$$\sqrt{k} \sum_{t=1}^{k} \Delta_t \xrightarrow{d} N(\mathbf{0}, \mathbf{G}^{-1}\Sigma\mathbf{G}^{-1}).$$

Lemma C.3. *Suppose that Assumptions 1, 2 (with* $p > 2$*) and 3 hold. Then*

$$\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \mu_t^{(0)} \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{n^2}\mathbf{S}\right), \tag{C.4}$$

where

$$\mu_t^{(0)} := -\frac{1}{n} \sum_{j=1}^{n} \nabla g_j(x^*; \zeta_{j,t}), \quad \mathbf{S} = \mathrm{Cov}\left(\sum_{j=1}^{n} \nabla g_j(x^*; \zeta_j)\right).$$

Proof. We show (C.4) by [1, Lemma 3.3.1].

Denote

$$\xi_{k,t} = \frac{\mu_t^{(0)}}{\sqrt{k}}, \quad \mathbf{S}_{k,t} = \mathbb{E}\left[\xi_{k,t}\xi_{k,t}^\intercal\right], \quad \mathbf{S}_k = \sum_{t=0}^{k-1} \mathbf{S}_{k,t},$$
$$\mathbf{R}_{k,t} = \mathbb{E}\left[\xi_{k,t}\xi_{k,t}^\intercal \middle| \xi_{k,0}, \cdots, \xi_{k,t-1}\right].$$

Then Lemma C.3 falls into the setting of [1, Lemma 3.3.1]. We just need to verify the conditions of [1, Lemma 3.3.1].

Since $\{\mu_t^{(0)}\}$ is a martingale difference sequence,

$$\mathbb{E}\left[\xi_{k,t} \middle| \xi_{k,0}, \cdots, \xi_{k,t-1}\right] = 0,$$

which implies the condition (3.3.1) of [1, Lemma 3.3.1].

Next, we verify the conditions (3.3.2)-(3.3.3) of [1, Lemma 3.3.1]. By the definition of $\xi_{k,t}$,

$$\begin{aligned}\mathbb{E}\left[\|\xi_{k,t}\|^p\right] &= \mathbb{E}\left[\left\|\frac{\frac{1}{n}\sum_{j=1}^n \nabla g_j(x^*;\zeta_{j,t})}{\sqrt{k}}\right\|^p\right] \\ &\leq \frac{\sum_{j=1}^n \mathbb{E}\left[\|\nabla g_j(x^*;\zeta_{j,t})\|^p\right]}{nk^{p/2}} \leq \frac{c_f^{p/2}}{k^{p/2}}, \end{aligned} \tag{C.5}$$

where the last inequality follows from Assumption 2. Then,

$$\sup_{k\geq 1}\sum_{t=0}^{k-1}\mathbb{E}\left[\|\xi_{k,t}\|^2\right] \leq \sup_{k\geq 1}\sum_{t=0}^{k-1}\left(\mathbb{E}\left[\|\xi_{k,t}\|^p\right]\right)^{2/p} \leq \sup_{k\geq 1}\frac{kc_f}{k} = c_f. \tag{C.6}$$

Note that $\mathbf{S}_{k,t} = \frac{1}{k}\frac{1}{n^2}\mathrm{Cov}\left(\sum_{j=1}^n \nabla g_j(x^*;\zeta_j)\right) = \frac{1}{k}\frac{1}{n^2}\mathbf{S}$,

$$\mathbf{S}_k = \sum_{t=0}^{k-1}\mathbf{S}_{k,t} = \frac{1}{n^2}\mathbf{S}. \tag{C.7}$$

(C.6) and (C.7) imply the condition (3.3.2) of [1, Lemma 3.3.1]. Moreover, the fact $\mathbf{R}_{k,t} = \mathbf{S}_{k,t}$ almost surely implies the condition (3.3.3) of [1, Lemma 3.3.1] directly.

It is left to verify the condition (3.3.4) of [1, Lemma 3.3.1]. For any $\delta > 0$, by the Hölder inequality,

$$\begin{aligned}\mathbb{E}\left[\|\xi_{k,t}\|^2\mathbf{1}_{\{\|\xi_{k,t}\|\geq\delta\}}\right] &\leq \left(\mathbb{E}\left[\|\xi_{k,t}\|^{2(p/2)}\right]\right)^{2/p}\left(\mathbb{E}\left[1^q_{\{\|\xi_{k,t}\|>\delta\}}\right]\right)^{1/q} \\ &= \left(\mathbb{E}\left[\|\xi_{k,t}\|^p\right]\right)^{2/p}\mathrm{P}^{1/q}\left(\|\xi_{k,t}\|>\delta\right) \\ &\leq \left(\mathbb{E}\left[\|\xi_{k,t}\|^p\right]\right)^{2/p}\left(\frac{\mathbb{E}\left[\|\xi_{k,t}\|\right]}{\delta}\right)^{1/q} \\ &\leq \frac{c_f^{1+1/(2q)}}{k^{1+1/(2q)}\delta^{1/q}},\end{aligned}$$

where p is defined in Assumption 2, q is the constant satisfying $2/p + 1/q = 1$, the second inequality follows from Markov inequality, the third inequality follows from (C.5). Then

$$\lim_{k\to\infty}\sum_{t=1}^{k}\mathbb{E}\left[\|\xi_{k,t}\|^2\mathbf{1}_{\{\|\xi_{k,t}\|\geq\delta\}}\right] \leq \lim_{k\to\infty}\frac{kc_f^{1+1/(2q)}}{k^{1+1/(2q)}\delta^{1/q}} = 0,$$

which implies the condition (3.3.4) of [1, Lemma 3.3.1].

Summarizing above, all the conditions of [1, Lemma 3.3.1] hold, then

$$\frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}\mu_t^{(0)} \xrightarrow{d} N\left(\mathbf{0}, \frac{1}{n^2}\mathbf{S}\right).$$

The proof is complete. □

Lemma C.4. *Suppose that Assumptions 1, 2 (with $p > 2$) and 3 hold. Then $\bar{x}_k$ converges to x^* almost surely.*

Proof. By the recursion $\bar{x}_{k+1} = \bar{x}_k - \alpha_k \bar{y}_k$,

$$\begin{aligned}\|\bar{x}_{k+1} - x^*\|^2 &= \|\bar{x}_k - x^*\|^2 - 2\alpha_k \langle \bar{x}_k - x^*, 1/n\nabla f(\bar{x}_k)\rangle \\ &\quad + 2\alpha_k \langle \bar{x}_k - x^*, 1/n\nabla f(\bar{x}_k) - \bar{y}_k\rangle + \|\alpha_k \bar{y}_k\|^2 \\ &\le \left(1 - \frac{\mu}{n}\alpha_k\right)\|\bar{x}_k - x^*\|^2 + 2\alpha_k \langle \bar{x}_k - x^*, 1/n\nabla f(\bar{x}_k) - \bar{y}_k\rangle + \|\alpha_k \bar{y}_k\|^2, \end{aligned} \tag{C.8}$$

where the inequality follows from the strongly convexity of $f(\cdot)$. Taking conditional expectation on both sides of (C.8),

$$\begin{aligned}\mathbb{E}\left[\|\bar{x}_{k+1} - x^*\|^2 \middle| \mathcal{F}_k\right] &\le \|\bar{x}_k - x^*\|^2 + 2\alpha_k \left\langle \bar{x}_k - x^*, 1/n\nabla f(\bar{x}_k) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k})\right\rangle \\ &\quad + \mathbb{E}\left[\|\alpha_k \bar{y}_k\|^2 \middle| \mathcal{F}_k\right] \\ &\le \left(1 + \alpha_k^2\right)\|\bar{x}_k - x^*\|^2 + \left\|1/n\nabla f(\bar{x}_k) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k})\right\|^2 \\ &\quad + \mathbb{E}\left[\|\alpha_k \bar{y}_k\|^2 \middle| \mathcal{F}_k\right], \end{aligned} \tag{C.9}$$

where the first inequality follows from the fact

$$\mathbb{E}\left[\bar{y}_k \middle| \mathcal{F}_k\right] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \nabla g(x_{j,k}; \zeta_{j,k}) \middle| \mathcal{F}_k\right] = \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k})$$

and the second inequality follows from the fact $2\langle a, b\rangle \le 2\|a\|\|b\| \le \|a\|^2 + \|b\|^2$. Denote

$$\begin{aligned} v_k &= \|\bar{x}_k - x^*\|^2, \quad a_k = \alpha_k^2, \\ b_k &= \left\|1/n\nabla f(\bar{x}_k) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,k})\right\|^2 + \mathbb{E}\left[\|\alpha_k \bar{y}_k\|^2 \middle| \mathcal{F}_k\right]. \end{aligned}$$

Then (C.9) can be rewritten as

$$\mathbb{E}[v_{k+1}|\mathcal{F}_k] \le (1 + a_k)v_k + b_k,$$

which is in the form of (C.1) in Lemma C.1.

By the Lipschitz continuity of $\nabla f_j(\cdot)$,

$$\sum_{t=0}^\infty \mathbb{E}\left[\left\|1/n\nabla f(\bar{x}_t) - \frac{1}{n}\sum_{j=1}^n \nabla f_j(x_{j,t})\right\|^2\right] \le \frac{L^2}{n}\sum_{t=0}^\infty \mathbb{E}\left[\|x_t - \mathbf{1}\otimes \bar{x}_t\|^2\right] \le \frac{L^2\bar{c}}{n}\sum_{t=0}^\infty \alpha_t^2 < \infty,$$

and by the Lipschitz continuity of $\nabla g_j(\cdot;\zeta_j)$,

$$
\begin{aligned}
&\sum_{t=0}^{\infty}\mathbb{E}\left[\|\alpha_t\bar{y}_t\|^2\right]\\
&=\sum_{t=0}^{\infty}\alpha_t^2\mathbb{E}\left[\left\|1/n\sum_{j=1}^{n}(\nabla g_j(x_{j,t};\zeta_{j,t})-\nabla g_j(x^*;\zeta_{j,t}))+1/n\sum_{j=1}^{n}\nabla g_j(x^*;\zeta_{j,t})\right\|^2\right]\\
&\leq\sum_{t=0}^{\infty}\alpha_t^2\left(2L^2/n\mathbb{E}\left[\|x_t-\mathbf{1}\otimes x^*\|^2\right]+2c_f\right)\\
&=\sum_{t=0}^{\infty}\mathcal{O}(\alpha_t^2)<\infty,
\end{aligned}
$$

where the second equality follows from Lemma 2.

Monotone convergence theorem implies

$$
\sum_{t=0}^{\infty}\left\|1/n\nabla f(\bar{x}_t)-\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(x_{j,t})\right\|^2<\infty,\quad \sum_{t=0}^{\infty}\mathbb{E}\left[\|\alpha_k\bar{y}_k\|^2\,\middle|\mathcal{F}_k\right]<\infty. \tag{C.10}
$$

Then $\sum_{k=1}^{\infty}b_k<\infty$ almost surely. By Lemma C.1, $\|\bar{x}_k-x^*\|^2$ converges to some random variable.

Combine Theorem 3 with Fatou lemma,

$$
\mathbb{E}\left[\liminf_{k\to\infty}\|\bar{x}_k-x^*\|^2\right]\leq\liminf_{k\to\infty}\mathbb{E}\left[\|\bar{x}_k-x^*\|^2\right]=0.
$$

Then $\bar{x}_k$ converges to x^* almost surely. The proof is complete. □

References

[1] Han. Fu Chen. *Stochastic approximation and its applications*, volume 64. Kluwer Academic Publishers, New York, 2006.

[2] John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21 – 48, 2021.

[3] Zhengyan Lin and Zhidong Bai. *Probability inequalities*. Springer Berlin, Heidelberg, 2011.

[4] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[5] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pages 233–257, 1971.