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

A large-scale stochastic gradient descent algorithm over a graphon

Yan Chen,[1] and Tao Li[*,2]

Abstract— **We study the large-scale stochastic gradient descent algorithm over a graphon with a continuum of nodes, which is regarded as the limit of the distributed networked optimization as the number of nodes goes to infinity. Each node has a private local cost function. The global cost function, which all nodes cooperatively minimize, is the integral of the local cost functions on the node set. We propose a stochastic gradient descent algorithm evolving as a graphon particle system, where each node heterogeneously interacts with others through a coupled mean field term. It is proved that if the graphon is connected, then by properly choosing the algorithm gains, all nodes' states achieve consensus uniformly in mean square. Furthermore, if the local cost functions are strongly convex, then all nodes' states converge uniformly to the minimizer of the global cost function in mean square.**

I. Introduction

In a distributed optimization problem over a network, all nodes cooperatively optimize a global cost function which is the sum of all local cost functions, and each node only knows its own local cost function. The distributed optimization algorithms involving information exchanging among nodes over a large-scale network can be found applications in distributed machine learning ([1]), multi-agent target tracking ([2]), distributed resource allocation ([3]-[4]), and so on. The dimensions of these algorithms explode as the number of nodes increases, and it is of interest to investigate the limiting case as the number of nodes tends to infinity. In fact, games and optimal control problems with a continuum of individuals have been studied intensively in the field called mean field games, which was pioneered independently by Huang, Malhamé and Caines ([5]) and Lasry and Lions ([6]), repectively. They attempt to understand the behaviors of the limiting systems of the dynamic games with a large number of individuals. In the past decades, there has been an increasing intention in mean field games and their applications ([7]-[11]). Motivated by the distributed optimization over large-scale networks and the developing theory of mean-field control and games, we investigate the limiting model of the distributed optimization problem as the number of nodes tends to infinity, that is, the distributed optimization problem over a graphon with a continuum of nodes.

Let $[0,1]$ be the set of a continuum of nodes, each element of which corresponds to a node. The connecting structure among nodes is given by the graphon A, which is a symmetric measurable function from $[0,1]\times[0,1]$ to $[0,1]$ ([12]). Any node $p\in[0,1]$ has a private local cost function $V(p,x):[0,1]\times\mathbb{R}^n\to\mathbb{R}$, which is strongly convex and continuously differentiable with respect to $x\in\mathbb{R}^n$ and is integrable with respect to $p\in[0,1]$. The objective of all nodes is to cooperatively solve the optimization problem

$$\min_{x\in\mathbb{R}^n} V(x) \triangleq \int_{[0,1]} V(p,x)dp. \tag{1}$$

Denote the unique minimizer of $V(x)$ by x^*.

In the distributed optimization over a network with finite nodes, all nodes interact through the underlying network. The interactions among nodes depend on their labels and so are heterogenous. In the graphon mean field theory, the concept of graph limit is introduced into the mean field theory, which provides a powerful tool for modeling the heterogeneous interactions among a large number of individuals ([13]-[16]). Representing the heterogeneous interactions among nodes in terms of the coupled mean field terms based on the graphon, we propose the following distributed stochastic gradient descent algorithm for the problem (1): given the initial states $\{x_p(0),p\in[0,1]\}$, for any node $p\in[0,1]$,

$$\begin{aligned} dx_p(t) =&\beta(t)\int_{\mathbb{R}^n\times[0,1]} A(p,q)(x-x_p(t))\mu_t(dx,dq)dt \\ &-\alpha(t)\nabla_x V\left(p,x_p(t)\right)dt-\alpha(t)\Sigma dw_p(t), \end{aligned} \tag{2}$$

where $x_p(t)\in\mathbb{R}^n$ is the state of node p at time t, representing its local estimate of x^*; $\nabla_x V(p,x_p(t))\in\mathbb{R}^n$ is the gradient value of the local cost function at the state $x_p(t)$; $\int_{\mathbb{R}^n\times[0,1]} A(p,q)(x-x_p(t))\mu_t(dx,dq)$ is the coupled mean field term based on the graphon A. Let $(\Omega,\mathcal{F},\mathcal{P})$ be a complete probability space with a family of non-decreasing σ-algebras $\{\mathcal{F}_t,\ t\geqslant 0\}\subseteq\mathcal{F}$. For any $t\geqslant 0$, $\mu_t(dx,dq)$ is a distribution on $\mathbb{R}^n\times[0,1]$ and satisfies: (i) the marginal distribution $\mu_t(dq)$ is always the uniform distribution on $[0,1]$, that is, $\mu_t(dq)=dq,\ \forall\ t\geqslant 0$; (ii) given $q\in[0,1]$, the conditional distribution $\mu_t(dx|q)$ is the distribution of $x_q(t)$. Here, $\{w_p(t),\ t\geqslant 0,\ p\in[0,1]\}$ is a family of independent n-dimensional standard Brownian motions, $x_p(0)$ is independent of $\{w_p(t),\ t\geqslant 0\}$ and adapted to $\mathcal{F}_0$, $w_p(t)$ is adapted to $\mathcal{F}_t,\ \forall\ t\geqslant 0,\ p\in[0,1]$, $\alpha(t)$ and $\beta(t)$ are time-varying algorithm gains and $\Sigma\in\mathbb{R}^{n\times n}$.

Denote the conditional distribution $\mu_t(dx|q)$ by $\mu_{t,q}(dx)$. Then we have $\mu_t(dx,dq)=\mu_{t,q}(dx)dq$. Therefore, (2) can

[*] Author for correspondence.

This work was supported in part by the National Natural Science Foundation of China under Grant 62261136550 and in part by the Basic Research Project of Shanghai Science and Technology Commission under Grant 20JC1414000.

[1] Yan Chen is with the School of Mathematical Sciences, East China Normal University, Shanghai 200241, China. Email: YanChen@stu.ecnu.edu.cn

[2] Tao Li is with the Key Laboratory of Pure Mathematical Practice, School of Mathematical Sciences, East China Normal University, Shanghai 200241, China. Email: tli@math.ecnu.edu.cn

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

be written as

$$\begin{aligned} dx_p(t) =&\beta(t)\int_{[0,1]}\left(\int_{\mathbb{R}^n} A(p,q)(x-x_p(t))\mu_{t,q}(dx)\right)dqdt \\ &- \alpha(t)\nabla_x V\left(p,x_p(t)\right)dt - \alpha(t)\Sigma dw_p(t). \end{aligned} \tag{3}$$

By using spatial and temporal discretion, we can show how (1) and (2) are related to the distributed optimization over the network with finite nodes. For any given positive integer N, we define $V^N(p,x) = V(\frac{i}{N},x)$, $p \in \left(\frac{i-1}{N},\frac{i}{N}\right]$, $i = 1,\cdots,N$. Then, $\int_{[0,1]} V^N(p,x)dp$ approximates $\int_{[0,1]} V(p,x)dp$ if N tends to infinity. Define $v_{N,i}(x) = V^N(\frac{i}{N},x)$, $i = 1,\cdots,N$. Then one obtains a distributed optimization problem over the network with N nodes:

$$\min_{x\in\mathbb{R}^n}\int_{[0,1]} V^N(p,x)dp = \min_{x\in\mathbb{R}^n}\frac{1}{N}\sum_{i=1}^N v_{N,i}(x).$$

Define a step graphon as $A^N(p,q) = A\left(\frac{i}{N},\frac{j}{N}\right)$, $p \in \left(\frac{i-1}{N},\frac{i}{N}\right]$, $q \in \left(\frac{j-1}{N},\frac{j}{N}\right]$, $i,\ j = 1,\cdots,N$, and then A^N approximates A if N tends to infinity ([13]). Define $x_p^N(t) = x_{\frac{i}{N}}(t)$ and $w_p^N(t) = w_{\frac{i}{N}}(t)$, $p \in \left(\frac{i-1}{N},\frac{i}{N}\right]$, $i = 1,\cdots,N$. Let $\mu_t^N(dx,dq)$ be a distribution on $\mathbb{R}^n \times [0,1]$ satisfying: (i) the marginal distribution $\mu_t^N(dq)$ is always the uniform distribution over $[0,1]$, that is, $\mu_t^N(dq) = dq$, $\forall\ t \geqslant 0$; (ii) for any $j = 1,\cdots,N$, given $q \in \left(\frac{j-1}{N},\frac{j}{N}\right]$, the conditional distribution $\mu_t^N\left(dx|q\right) = \delta_{x^N_{\frac{j}{N}}(t)}(dx)$, where $\delta_{x^N_{\frac{j}{N}}(t)}(dx)$ is the Dirac measure at $x^N_{\frac{j}{N}}(t)$. Therefore, $\mu_t^N(dx,dp)$ approximates $\mu_t(dx,dp)$ if N tends to infinity, which together with (2) yields the following system: for any $p \in (0,1]$,

$$\begin{aligned} dx_p^N(t) =&\beta(t)\int_{\mathbb{R}^n\times[0,1]} A^N(p,q)(x-x_p^N(t))\mu_t^N(dx,dq)dt \\ &- \alpha(t)\nabla_x V^N\left(p,x_p^N(t)\right)dt - \alpha(t)\Sigma dw_p^N(t). \end{aligned} \tag{4}$$

By the graph limit theory and the mean field theory, the above system approximates (3) if N tends to infinity. In particular, take $p = \frac{i}{N}$ in (4) respectively and denote $x_{N,i}(t) = x^N_{\frac{i}{N}}(t)$, $w_{N,i}(t) = w^N_{\frac{i}{N}}(t)$, $i = 1,\cdots,N$, and $a_{N,ij} = A^N\left(\frac{i}{N},\frac{j}{N}\right)$, $i,\ j = 1,\cdots,N$. Then we have the N particle systems:

$$\begin{aligned} &dx_{N,i}(t) \\ =&\beta(t)\sum_{j=1}^N\int_{\mathbb{R}^n\times\left(\frac{j-1}{N},\frac{j}{N}\right]} a_{N,ij}(x-x_{N,i}(t))\mu_t^N(dx,dq)dt \\ &- \alpha(t)\nabla_x v_{N,i}(x_{N,i}(t))dt - \alpha(t)\Sigma dw_{N,i}(t) \\ =&\beta(t)\sum_{j=1}^N\int_{\left(\frac{j-1}{N},\frac{j}{N}\right]}\left(\int_{\mathbb{R}^n} a_{N,ij}(x-x_{N,i}(t))\right. \\ &\left.\mu_t^N\left(dx|q\right)\right)dqdt - \alpha(t)\nabla_x v_{N,i}(x_{N,i}(t))dt \\ &- \alpha(t)\Sigma dw_{N,i}(t) \\ =&\beta(t)\sum_{j=1}^N\int_{\left(\frac{j-1}{N},\frac{j}{N}\right]}\left(\int_{\mathbb{R}^n} a_{N,ij}(x-x_{N,i}(t))\right. \\ &\left.\delta_{x_{N,j}(t)}(dx)\right)dqdt - \alpha(t)\nabla_x v_{N,i}(x_{N,i}(t))dt \\ &- \alpha(t)\Sigma dw_{N,i}(t) \\ =&\beta(t)\sum_{j=1}^N\int_{\left(\frac{j-1}{N},\frac{j}{N}\right]} a_{N,ij}(x_{N,j}(t)-x_{N,i}(t))dqdt \\ &- \alpha(t)\nabla_x v_{N,i}(x_{N,i}(t))dt - \alpha(t)\Sigma dw_{N,i}(t) \\ =&\beta(t)\frac{1}{N}\sum_{j=1}^N a_{N,ij}(x_{N,j}(t)-x_{N,i}(t))dt \\ &- \alpha(t)\nabla_x v_{N,i}(x_{N,i}(t))dt - \alpha(t)\Sigma dw_{N,i}(t), i = 1,\cdots,N. \end{aligned}$$

For a given sequence $0 \leqslant t_0 < t_1 < \cdots < t_k < \cdots$ in the time interval $[0,\infty)$, where $t_k \geqslant 0$ and $k = 0,\ 1,\ 2\cdots$, by [17], the Euler approximation of the above stochastic differential equation is given by

$$\begin{aligned} x_{N,i}(t_{k+1}) =&x_{N,i}(t_k) + \beta(t_k)(t_{k+1}-t_k)\frac{1}{N}\sum_{j=1}^N a_{N,ij} \\ &\times(x_{N,j}(t_k)-x_{N,i}(t_k)) - \alpha(t_k)(t_{k+1}-t_k) \\ &\times\left(\nabla_x v_{N,i}(x_{N,i}(t_k)) + \xi_{N,i}(t_k)\right), \end{aligned} \tag{5}$$

where $\xi_{N,i}(t_k) = \Sigma(w_{N,i}(t_{k+1}) - w_{N,i}(t_k))$ is an n-dimensional martingale difference sequence with zero mean and covariance matrix $(t_{k+1}-t_k)\Sigma\Sigma^T$. It can be verified that (5) is just the distributed optimization algorithm over the network with finite nodes in [18]-[21].

If $\alpha(t) = 0$ in (3), then $x_q(t)$ degenerates to a deterministic progress and $\mu_{t,q}(dx)$ degenerates to $\delta_{x_q(t)}(dx)$. Then, one obtains a special case of (3) given by

$$\begin{aligned} &dx_p(t) \\ =&\beta(t)\int_{[0,1]}\left(\int_{\mathbb{R}^n} A(p,q)(x-x_p(t))\delta_{x_q(t)}(dx)\right)dqdt \\ =&\beta(t)\int_{[0,1]} A(p,q)(x_q(t)-x_p(t))dqdt, \end{aligned} \tag{6}$$

which is called the first-order consensus system ([22]-[24]).

If the distribution of $x_p(0)$, $V(p,x)$, $w_p(t)$, $A(p,q)$ and $\mu_{t,p}(dx)$ do not depend on the label p in (3), and are denoted by μ_0, $V(x)$, $w(t)$, A_q, $\mu_t(dx)$, respectively, then the system (3) degenerates to

$$\begin{aligned} dx(t) =&\beta(t)\left(\int_{[0,1]} A_q dq\right)\left(\int_{\mathbb{R}^n}(x-x(t))\mu_t(dx)\right)dt \\ &- \alpha(t)\nabla_x V(x(t))dt - \alpha(t)\Sigma dw(t), \end{aligned}$$

in the sense of weak solution, which is the classical Mckean-Vlasov equation ([25]-[26]).

In fact, the algorithm (3) belongs to a class of particle systems with heterogeneous interactions: graphon particle systems, for which fruitful results have been achieved. Works ([13]-[14]) have focused on the existence and uniqueness of the solution for different graphon particle systems and the convergence of finite particle systems to graphon particle systems. Only few works ([15]-[16]) are concerned with the asymptotic properties of graphon particle systems. Bayraktar and Wu ([16]) showed that the distribution of each node's state and the integral of the distributions on the node set converge to the limiting distribution and the integral of the limiting distributions on the node set respectively.

Note that all aforementioned works on graphon particle systems do not reveal the relation between the limiting distribution and system dynamics. However, for many practical applications, people are more interested in how the limiting distribution is associated with the system dynamics. In particular, for the problem (1) and the algorithm (3), people expect to figure out whether the states $\{x_p(t),\ p \in [0,1],\ t \geqslant 0\}$ of the system (3) converge to the minimizer of the global cost function under some proper assumptions. However, all existing works are unable to address the issue.

In this paper, we prove that if the graphon is connected and the local cost functions are strongly convex, then by properly choosing algorithm gains, the states $\{x_p(t),\ p \in [0,1],\ t \geqslant 0\}$ of the system (3) converge to the minimizer of the global cost function in mean square. Different from Bayraktar et al. ([16]), we weaken assumptions on local cost functions and yield stronger results. Bayraktar et al. ([16]) assumed that the dissipativity of the drift term is strictly twice greater than the Lipschitz constant of the interaction term. For the system (2), this assumption is equivalent to the strong convexity constant of the local cost functions being greater than 2, which is not reasonable for distributed optimization problems. In this paper, the local cost functions are only assumed to be strongly convex and there is no further requirement on the strong convexity constant. Bayraktar et al. ([16]) proved that the all nodes' states converge in distribution, while we prove the convergence in mean square. Bayraktar et al. ([16]) proved the existence of the limiting distributions of the nodes' states. We not only prove the existence of the limiting distribution but also reveal that the limiting distribution is the Dirac distribution at the minimizer of the global cost function.

Compared with the time-invariant graphon particle system in [16], the system (3) is time-varying due to the time-varying algorithm gains introduced. The introducing of time-varying algorithm gains removes the requirement on the strong convexity constant of the local cost functions, while it poses difficulties in the uniform boundedness of the second moments of all nodes' states. We prove that the second moments of all nodes' states are uniformly bounded in two steps. At first, the uniform boundedness of $\int_{[0,1]} E\big[\|x_p(t)\|^2\big]dp$ is proved by using strictly positive algebraic connectivity of the graphon. Then, by properly choosing the algorithm gains, we prove that the second moments of all nodes' states are uniformly bounded.

We prove that if the graphon is connected, then all nodes' states in the system (3) achieve consensus in mean square, that is, $\lim_{t\to\infty} \sup_{p\in[0,1]} E\big[\|x_p - \int_{[0,1]} x_q(t)dq\|^2\big] = 0$. To prove this, we obtain that $\lim_{t\to\infty} \int_{[0,1]} E\big[\|x_p - \int_{[0,1]} x_q(t)dq\|^2\big]dp = 0$ by the connectivity of the graphon and the Lyapunov method firstly. We qualify how the convergence rate of $\lim_{t\to\infty} \int_{[0,1]} E\big[\|x_p - \int_{[0,1]} x_q(t)dq\|^2\big]dp = 0$ relates to the system dynamics (3), especially, the algebraic connectivity of the graphon. Then, by exploiting the uniform boundness of the second moments of all nodes' states and combining $\lim_{t\to\infty} \int_{[0,1]} E[\|x_p - \int_{[0,1]} x_q(t)dq\|^2]dp = 0$ with the Lyapunov method in the consensus error of each node, we prove that all nodes' states in the system (3) achieve consensus uniformly in mean square, which in turn derives that all nodes' states converge to the minimizer of the global cost function uniformly in mean square.

The remainder of this paper is organized as follows. In Section II, we prove the convergence of the algorithm. In Section III, conclusions are given. In Appendix, we provide the definitions of the connectivity and the algebraic connectivity of a graphon. Due to the space limitation, proofs of some lemmas and theorems are omitted.

The following notations will be used throughout this paper. Denote the set of all real numbers by $\mathbb{R}$. Denote the n-dimensional Euclidean space by $\mathbb{R}^n$ and the Euclidean norm by $\|\cdot\|$. For a given matrix $A \in \mathbb{R}^{n\times n}$, $\mathrm{Tr}(A)$ denotes the trace of A. For a given vector $x \in \mathbb{R}^n$, x^T denotes the transpose of x. Denote $L^2([0,1],\mathbb{R}^n) = \{f : [0,1] \to \mathbb{R}^n,\ \int_{[0,1]} \|f(x)\|^2 dx < \infty\}$. Denote the set of all bounded linear operators from $L^2([0,1],\mathbb{R}^n)$ to $L^2([0,1],\ \mathbb{R}^n)$ by $\mathcal{L}\big(L^2([0,1],\ \mathbb{R}^n)\big)$. Denote the inner product on $L^2([0,1],\ \mathbb{R}^n)$ by $\langle\cdot,\cdot\rangle_{L^2([0,1],\mathbb{R}^n)}$, that is, for any given $f,\ g \in L^2([0,1],\ \mathbb{R}^n)$, $\langle f,g\rangle_{L^2([0,1],\ \mathbb{R}^n)} \triangleq \int_{[0,1]} f^T(x)g(x)dx$. For a given function $f : F \to \mathbb{R}$, $\mathrm{supp}(f) = \{x \in F : f(x) \neq 0\}$ denotes the support set of f. For a given random variable $X \in \mathbb{R}^n$, denote the mathematical expectation of X by $E[X]$. For a given measurable space $(F,\ \mathscr{G})$ and $x \in F$, where $\mathscr{G}$ is a σ-algebra in F, Dirac measure δ_x at x is the measure defined by $\delta_x(A) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases},\ \forall\ A \in \mathscr{G}$.

II. Convergence of the Algorithm

We suppose that for any given $T > 0$, there exists a unique solution $\big\{x_p(t),\ \mu_{t,p},\ t \in [0,T],\ p \in [0,1]\big\}$ for the graphon particle system (3), satisfying $\sup_{p\in[0,1]} \sup_{t\in[0,T]} E[\|x_p(t)\|^2] < \infty$, where $\mu_{t,p}$ is the distribution of $x_p(t)$. In this section, we prove the convergence of the stochastic gradient descent algorithm (3) by the asymptotic property of the solution for the graphon particle system. Firstly, we prove that the variance of each node's state converges to zero uniformly by properly choosing algorithm gains in (3). Secondly, combining the connectivity of the graphon and the uniform boundness of second moments of the states, we prove that all nodes' states achieve the consensus uniformly in mean square. Finally, the uniform convergence of all nodes' states to the minimizer of the global cost function is given. We make the following assumptions on the graphon particle system (3).

Assumption 2.1: The graphon A is connected.

Assumption 2.2: (i) $\sup_{p\in[0,1]} E\left[\|x_p(0)\|^2\right] < \infty$;

(ii) There exists a constant $\kappa > 0$ such that $\|\nabla_x V(p,x) - \nabla_x V(p,\widetilde{x})\| \leqslant \kappa \|x-\widetilde{x}\|,\ \forall\ x,\ \widetilde{x}\in\mathbb{R}^n,\ p\in[0,1]$;

(iii) For any $p \in [0,1],\ x\in\mathbb{R}^n$, the local cost function is continuously differentiable and uniformly strongly convex with respect to x, that is, there exists $\kappa_2 > 0$ such that $(x-\widetilde{x})^T\left(\nabla_x V(p,x) - \nabla_x V(p,\widetilde{x})\right) \geqslant \kappa_2\|x-\widetilde{x}\|^2,\ \forall\ x,\ \widetilde{x}\ \in\mathbb{R}^n,\ p\in[0,1]$.

Assumption 2.3: The time-varying algorithm gains satisfy $\alpha(t)>0,\ \beta(t)>0,\ \int_0^\infty \beta(s)ds=\infty,\ \int_0^\infty\alpha^2(s)ds<\infty$, $\lim_{t\to\infty}\beta(t)=0$ and $\lim_{t\to\infty}\frac{\alpha(t)}{\beta(t)}=0$.

Denote $\zeta=\sup_{p\in[0,1]}E\left[\|x_p(0)\|^2\right]$.

In the following lemma, we present the convergence result for the variance of each node's state.

Lemma 2.1: For the problem (1) and the algorithm (3), if Assumptions 2.1-2.3 hold, then

$$\lim_{t\to\infty}\sup_{p\in[0,1]}E\left[\|x_p(t)-E[x_p(t)]\|^2\right]=0.$$

Proof: Noting that $\mu_{t,p}(dx)$ is the distribution of $x_p(t)$ in (3), the graphon particle system (3) can be written as

$$\begin{aligned}dx_p(t)=&\beta(t)\int_{[0,1]}A(p,q)\left(E[x_q(t)]-x_p(t)\right)dqdt\\&-\alpha(t)\nabla_xV\left(p,x_p(t)\right)dt-\alpha(t)\Sigma dw_p(t).\end{aligned}\quad(7)$$

By $\int_0^\infty\alpha^2(s)ds<\infty$ in Assumption 2.3 and [27], we have $E\Big[\int_0^t\alpha(s)\Sigma dw_p(s)\Big]=0,\forall\ p\in[0,1]$. This together with (7) leads to

$$\begin{aligned}E[x_p(t)]=&E[x_p(0)]+\int_0^t\Big[\beta(s)\int_{[0,1]}A(p,q)\Big(E[x_q(s)]\\&-E[x_p(s)]\Big)dq-\alpha(s)E\left[\nabla_xV(p,x_p(s))\right]\Big]ds.\end{aligned}$$

This can be written as

$$\begin{aligned}\frac{dE[x_p(t)]}{dt}=&\beta(t)\int_{[0,1]}A(p,q)\left(E[x_q(t)]-E[x_p(t)]\right)dq\\&-\alpha(t)E\left[\nabla_xV\left(p,x_p(t)\right)\right].\end{aligned}$$

Denote $S_p(t)=\|x_p(t)-E[x_p(t)]\|^2$. By the above equation, (7) and Itô's formula, we have

$$\begin{aligned}&S_p(t)-S_p(0)\\=&\int_0^t2\left(x_p(s)-E[x_p(s)]\right)^T\Big(\beta(s)\int_{[0,1]}A(p,q)dq)\\&\times\left(E\left[x_p(s)\right]-x_p(s)\right)-\alpha(s)\Big(\nabla_xV(p,x_p(s))\\&-E\left[\nabla_xV(p,x_p(s))\right]\Big)\Big)ds-\int_0^t2\alpha(s)\Big(x_p(s)\\&-E[x_p(s)]\Big)^T\Sigma dw_p(s)+\int_0^t\alpha^2(s)\operatorname{Tr}(\Sigma^T\Sigma)ds.\end{aligned}\quad(8)$$

By Assumption 2.3, we know that there exists a constant $\alpha_1>0$ such that $\sup_{t\geqslant0}\alpha(t)\leqslant\alpha_1$. This together with C_2 inequality and Jensen inequality leads to

$$\begin{aligned}&E\left[\int_0^t\left\|2\alpha(s)\left(x_p(s)-E[x_p(s)]\right)^T\Sigma\right\|^2ds\right]\\\leqslant&4\alpha_1^2\|\Sigma\|^2E\left[\int_0^t\left(2\|x_p(s)\|^2+2\|E[x_p(s)]\|^2\right)ds\right]\\=&8\alpha_1^2\|\Sigma\|^2E\left[\int_0^t\|x_p(s)\|^2ds\right]\\\leqslant&8t\alpha_1^2\|\Sigma\|^2\sup_{0\leqslant s\leqslant t}E\left[\|x_p(s)\|^2\right]<\infty,\end{aligned}$$

then, by [27], we have $E\Big[\int_0^t2\alpha(s)\left(x_p(s)-E[x_p(s)]\right)^T\Sigma dw_p(s)\Big]=0$. Noticing that for the given p, the terms $\nabla_xV(p,E[x_p(t)])]$ and $E[\nabla_xV(p,x_p(t))]$ are deterministic, then

$$\begin{aligned}E\Big[&\left(x_p(t)-E[x_p(t)]\right)^T\Big(\nabla_xV(p,E[x_p(t)])\\&-E[\nabla_xV(p,x_p(t))]\Big)\Big]=0.\end{aligned}$$

Then, by (8) and Assumption 2.2, we have

$$\begin{aligned}&\frac{dE[S_p(t)]}{dt}\\=&E\Big[2\left(x_p(t)-E[x_p(t)]\right)^T\Big(\beta(t)\int_{[0,1]}A(p,q)dq\Big(E[x_p(t)]\\&-x_p(t)\Big)-\alpha(t)\left(\nabla_xV(p,x_p(t))-E[\nabla_xV(p,x_p(t))]\right)\Big)\Big]\\&+\alpha^2(t)\operatorname{Tr}(\Sigma^T\Sigma)\\=&2\beta(t)\int_{[0,1]}A(p,q)dqE\Big[\left(x_p(t)-E[x_p(t)]\right)^T\big(E[x_p(t)]\\&-x_p(t)\big)\Big]-2\alpha(t)E\Big[\left(x_p(t)-E[x_p(t)]\right)^T\\&\times\left(\nabla_xV(p,x_p(t))-E\left[\nabla_xV(p,x_p(t))\right]\right)\Big]\\&+\alpha^2(t)\operatorname{Tr}(\Sigma^T\Sigma)\\=&-2\beta(t)\int_{[0,1]}A(p,q)dqE[S_p(t)]-2\alpha(t)E\Big[\big(x_p(t)\\&-E[x_p(t)]\big)^T\left(\nabla_xV(p,x_p(t))-\nabla_xV(p,E[x_p(t)])\right)\Big]\\&-2\alpha(t)E\Big[\big(x_p(t)-E[x_p(t)]\big)^T\big(\nabla_xV(p,E[x_p(t)])\\&-E[\nabla_xV(p,x_p(t))]\big)\Big]+\alpha^2(t)\operatorname{Tr}(\Sigma^T\Sigma)\\\leqslant&-\left(2\beta(t)\int_{[0,1]}A(p,q)dq+2\alpha(t)\kappa_2\right)E[S_p(t)]\\&+\alpha^2(t)\operatorname{Tr}(\Sigma^T\Sigma)\\\leqslant&-\phi(t)E\left[S_p(t)\right]+\alpha^2(t)\operatorname{Tr}(\Sigma^T\Sigma),\end{aligned}$$

where $\phi(t)=2\beta(t)\inf_{p\in[0,1]}\int_{[0,1]}A(p,q)dq+2\alpha(t)\kappa_2$. By the above inequality and the comparison theorem ([28]), we have

$$E\left[S_p(t)\right]\leqslant e^{-\int_0^t\phi(s)ds}E[S_p(0)]$$

$$+\operatorname{Tr}(\Sigma^T\Sigma)\int_0^t e^{-\int_s^t \phi(s')ds'}\alpha^2(s)ds.$$

By Assumption 2.2, we have $E[S_p(0)] \leqslant E[\|x_p(0)\|^2] \leqslant \zeta$. This together with the above inequality gives

$$\begin{aligned}&\sup_{p\in[0,1]} E\left[S_p(t)\right]\\ \leqslant & e^{-\int_0^t \phi(s)ds}\zeta + \operatorname{Tr}(\Sigma^T\Sigma)\int_0^t e^{-\int_s^t \phi(s')ds'}\alpha^2(s)ds. \quad (9)\end{aligned}$$

It follows from Assumptions 2.1-2.3 that

$$\lim_{t\to\infty} e^{-\int_0^t \phi(s)ds}\zeta = 0. \quad (10)$$

For the second term on the right side of (9), by Assumptions 2.1-2.3 and L'Hospital's rule, we have

$$\begin{aligned}&\lim_{t\to\infty}\int_0^t e^{-\int_s^t \phi(s')ds'}\alpha^2(s)ds\\ =&\lim_{t\to\infty}\frac{\int_0^t e^{\int_0^s \phi(s')ds'}\alpha^2(s)ds}{e^{\int_0^t \phi(s)ds}}\\ =&\lim_{t\to\infty}\frac{\alpha^2(t)}{2\beta(t)\inf\limits_{p\in[0,1]}\int_{[0,1]}A(p,q)dq + 2\alpha(t)\kappa_2}\\ =&\lim_{t\to\infty}\frac{\alpha(t)\frac{\alpha(t)}{\beta(t)}}{\left(2\inf\limits_{p\in[0,1]}\int_{[0,1]}A(p,q)dq + 2\frac{\alpha(t)}{\beta(t)}\kappa_2\right)} = 0.\end{aligned}$$

Combining (9) with (10) and the above equality gives $\lim_{t\to\infty}\sup_{p\in[0,1]} E\left[\|x_p(t) - E[x_p(t)]\|^2\right] = 0$. ■

Lemma 2.2: For the problem (1) and the algorithm (3), if Assumptions 2.1-2.3 hold, then there exists $K \geqslant 0$ such that $\sup_{p\in[0,1],\ t\geqslant 0} E\left[\|x_p(t)\|^2\right] \leqslant K$.

The following lemma illustrates that all nodes' states achieve consensus uniformly in mean square.

Lemma 2.3: For the problem (1) and the algorithm (3), if Assumptions 2.1-2.3 hold, then

$$\begin{aligned}&\int_{[0,1]} E\left[\left\|x_p(t) - \int_{[0,1]} x_q(t)dq\right\|^2\right]dp\\ \leqslant & 6e^{-\int_0^t \phi(s)ds}\zeta + 3e^{-\lambda_2(\mathbb{L}_A)\int_0^t \beta(s)ds}\zeta + 6\operatorname{Tr}(\Sigma^T\Sigma)\\ &\times\int_0^t e^{-\int_s^t \phi(s')ds'}\alpha^2(s)ds + 3\int_0^t\Bigg(\left(8\sigma_v K + 8C_v K^{\frac{1}{2}}\right)\\ &\alpha(s)e^{-2\lambda_2(\mathbb{L}_A)\int_s^t \beta(s')ds'}\Bigg)ds\end{aligned}$$

and

$$\lim_{t\to\infty}\sup_{p\in[0,1]} E\left[\left\|x_p(t) - \int_{[0,1]} x_q(t)dq\right\|^2\right] = 0,$$

where $\phi(t) = 2\alpha(t)\kappa_2 + 2\beta(t)\inf_{p\in[0,1]}\int_{[0,1]}A(p,q)dq$, K is given by Lemma 2.2, and $\lambda_2(\mathbb{L}_A)$ is the algebraic connectivity of the graphon A.

Below we will prove that the integral of the expected states on the node set converges to the minimizer of the global cost function. By Assumption 2.2, we know that $V(x)$ is strongly convex with respect to x. This together with the fact that for all $p \in [0,1]$, $\nabla_x V(p,x)$ is continuous with respect to $x \in \mathbb{R}^n$ leads to $\nabla_x V(x^*) = \int_{[0,1]}\nabla_x V(p,x^*)dp = 0$.

Lemma 2.4: For the problem (1) and the algorithm (3), if Assumptions 2.1-2.3 hold, then

$$\lim_{t\to\infty}\left\|\int_{[0,1]} E\left[x_p(t)\right]dp - x^*\right\|^2 = 0.$$

Combining the above lemmas, we obtain that all nodes' states converge to the minimizer of the global cost function uniformly in mean square.

Theorem 2.1: For the problem (1) and the algorithm (3), if Assumptions 2.1-2.3 hold, then

$$\lim_{t\to\infty}\sup_{p\in[0,1]} E\left[\|x_p(t) - x^*\|^2\right] = 0.$$

Proof: By Cauchy-Schwarz inequality and Hölder inequality, we have

$$\begin{aligned}&\sup_{p\in[0,1]}\left\|E[x_p(t)] - \int_{[0,1]} E[x_q(t)]dq\right\|^2\\ \leqslant & 3\sup_{p\in[0,1]} E\left[\|E[x_p(t)] - x_p(t)\|^2\right] + 3\sup_{p\in[0,1]} E\Bigg[\Bigg\|x_p(t)\\ &-\int_{[0,1]} x_q(t)dq\Bigg\|^2\Bigg] + 3E\Bigg[\Bigg\|\int_{[0,1]} x_q(t)dq\\ &-\int_{[0,1]} E[x_q(t)]dq\Bigg\|^2\Bigg]\\ \leqslant & 6\sup_{p\in[0,1]} E\left[\|E[x_p(t)] - x_p(t)\|^2\right] + 3\sup_{p\in[0,1]} E\Bigg[\Bigg\|x_p(t)\\ &-\int_{[0,1]} x_q(t)dq\Bigg\|^2\Bigg].\end{aligned}$$

By the above inequality, Lemma 2.1 and Lemma 2.3, we have

$$\lim_{t\to\infty}\sup_{p\in[0,1]}\left\|E[x_p(t)] - \int_{[0,1]} E[x_q(t)]dq\right\|^2 = 0. \quad (11)$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\sup_{p\in[0,1]} E\left[\|x_p(t) - x^*\|^2\right]\\ \leqslant & 3\sup_{p\in[0,1]} E\left[\|x_p(t) - E[x_p(t)]\|^2\right] + 3\sup_{p\in[0,1]}\Bigg\|E[x_p(t)]\\ &-\int_{[0,1]} E[x_q(t)]dq\Bigg\|^2 + 3\left\|\int_{[0,1]} E[x_q(t)]dq - x^*\right\|^2.\end{aligned}$$

This together with Assumptions 2.1-2.3, (11), Lemma 2.1 and Lemma 2.4 leads to $\lim_{t\to\infty}\sup_{p\in[0,1]} E\big[\|x_p(t) - x^*\|^2\big] = 0$. ■

Remark 2.1: The graphon particle system (3) is equivalent to the following system in the sense of weak solution: given the initial value $x(0)=x_P(0)$,

$$\begin{aligned} dx(t) =&\beta(t)\int_{\mathbb{R}^n\times[0,1]} A(P,q)(x-x(t))\mu_t(dx,dq)dt \\ &-\alpha(t)\nabla_x V(P,x(t))dt-\alpha(t)\Sigma dw(t), \end{aligned} \tag{12}$$

where P is a uniform random variable on $[0,1]$; for any $t\geqslant 0$, $\mu_t(dx,dq)$ is a distribution on $\mathbb{R}^n\times[0,1]$ and satisfies: (i) the marginal distribution $\mu_t(dq)$ is always the uniform distribution on $[0,1]$, that is, $\mu_t(dq)=dq,\ \forall\ t\geqslant 0$; (ii) the marginal distribution $\mu_t(dx)=\int_{[0,1]}\mu_t(dx|q)dq$ is the distribution of $x(t)$; $w(t)$ is an n-dimensional standard Brownian motion. From Theorem 2.1, we know that $\mu_t(dx|q)$ in (12) converges to $\delta_{x^*}(dx)$ uniformly. Then, the distribution $\mu_t(dx)$ converges to $\delta_{x^*}(dx)$.

III. Conclusion

In this paper, the large-scale stochastic gradient descent algorithm over the graphon has been studied. The evolution of the algorithm is characterized by a graphon particle system. By investigating the asymptotic property of the solution for the graphon particle system, we prove that all nodes' states achieve consensus uniformly in mean square if the graphon is connected and algorithm gains are chosen properly. Furthermore, we prove that if the local cost functions are strongly convex, then all nodes' states converge to the minimizer of the global cost function uniformly in mean square.

Appendix

For a given graphon W, the Graphon-Laplacian $\mathbb{L}_W\in\mathcal{L}\left(L^2\left([0,1],\ \mathbb{R}^n\right)\right)$ generated by W is given by: for any $z\in L^2([0,1],\ \mathbb{R}^n)$, $(\mathbb{L}_W z)(p)=\int_{[0,1]}W(p,q)(z(p)-z(q))dq,\ \forall\ p\in[0,1]$.

The algebraic connectivity of a graphon W is defined by $\lambda_2(\mathbb{L}_W)=\inf_{z\in\mathscr{C}}\frac{\langle\mathbb{L}_W z,z\rangle_{L^2([0,1],\ \mathbb{R}^n)}}{\langle z,z\rangle^2_{L^2([0,1],\ \mathbb{R}^n)}}\geqslant 0$, where $\mathscr{C}=\{z\in L^2\left([0,1],\ \mathbb{R}^n\right):\int_{[0,1]}z(\alpha)d\alpha=0\}$.

Definition A.1 ([23]): For a graphon W, if

(i) for any $p\in[0,1]$ and $q\in[0,1]\backslash\{p\}$, there exist integer $m\geqslant 1$ and a finite sequence $(l_k)_{1\leqslant k\leqslant m}\subset[0,1]$ satisfying $p=l_1,\ q=l_m$ and $l_{k+1}\in\operatorname{supp}\left(W\left(l_k,\cdot\right)\right)$, $\forall\ k\in\{1,\ldots,m-1\}$;

(ii) $\inf_{p\in[0,1]}\int_{[0,1]}W(p,q)\mathrm{d}q>0$,

then the graphon W is said to be connected.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, Fort Lauderdale, USA, Apr. 20-22, 2017, pp. 1273-1282.

[2] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification and tracking of targets in distributed sensor networks," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17-29, 2002.

[3] R. Madan and S. Lall, "Distributed algorithms for maximum lifetime routing in wireless sensor networks," *IEEE Trans. Wirel. Commun.*, vol. 5, no. 8, pp. 2185-2193, 2006.

[4] K. Wang, Z. Fu, Q. Xu, D. Chen, L. Wang, and W. Yu, "Distributed fixed step-size algorithm for dynamic economic dispatch with power flow limits," *Sci China Inf Sci*, vol. 64, pp. 1-13, 2021.

[5] M. Huang, R. P. Malhamé, and P. E. Caines, "On a class of large-scale cost-coupled Markov games with applications to decentralized power control," in *Proc. 43th IEEE Conf. Decision Control*, Nassau, Bahamas, Dec. 14-17, 2004, pp. 2830-2835.

[6] J. M. Lasry and P. L. Lions, "Jeux à champ moyen. I. Le cas stationnaire," *C. R. Math.*, vol. 343, no. 9, pp. 619-625, 2006.

[7] A. Lachapelle, J. Salomon, and G.Turinici, "Computation of mean field equilibria in economics," *Math. Models Meth. Appl. Sci.*, vol. 20, no. 4, pp. 567-588, 2010.

[8] D. Gomes, L. Lafleche, and L. Nurbekyan, "A mean-field game economic growth model," in *Proc. Amer. Control Conf.*, Boston, USA, Jul. 6-8, 2016, pp. 4693-4698.

[9] A. Bensoussan, K. C. J. Sung, S. C. P. Yam, and S. P. Yung, "Linear-quadratic mean field games," *J. Optim. Theory Appl.*, vol. 169, pp. 496-529, 2016.

[10] M. Bardi, "Explicit solutions of some linear-quadratic mean field games," *Netw. Heterog. Media*, vol. 7, no. 2, pp. 243-261, 2012.

[11] M. Bardi and F. S. Priuli, "Linear-quadratic N-person and mean-field games with ergodic cost," *SIAM J. Control Optim.*, vol. 52, no. 5, pp. 3022-3052, 2014.

[12] L. Lovasz, *Large Networks and Graph Limits*. Providence: American Mathematical Society, 2012.

[13] E. Bayraktar, S. Chakraborty and R. Wu. "Graphon mean field systems," arXiv: 2003.13180, 2020.

[14] G. Bet, F. Coppini, and F. R. Nardi, "Weakly interacting oscillators on dense random graphs," arXiv: 2006.07670, 2020.

[15] E. Bayraktar and R. Wu, "Graphon particle system:uniform-in-time concentration bounds," *Stoch. Process. Their Appl.*, vol. 156, pp. 196-225, 2023.

[16] E. Bayraktar and R. Wu, "Stationarity and uniform in time convergence for the graphon particle system," *Stoch. Process. Their Appl.*, vol. 150, pp. 532-568, 2022.

[17] E. Platen and N. Bruti-Liberati, *Numerical Solution of Stochastic Differential Equations with Jumps in Finance*. Berlin: Springer Science & Business Media, 2010.

[18] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48-61, 2009.

[19] K. Yuan, Q. Ling, and W. Yin," On the convergence of decentralized gradient descent," *SIAM J. Control Optim.*, vol. 26, no. 3, pp. 1835-1854, 2016.

[20] T. Li, K. Fu, X. Fu, and A. L. Fradkov, "Distributed stochastic optimization with Unbounded subgradients over randomly time-varying networks," arXiv: 2008.08796, 2020.

[21] B. Swenson, R. Murray , H. V. Poor, and S. Kar, " Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 14751-14812, 2022.

[22] B. E. Lee, "Consensus and voting on large graphs: an application of graph limit theory," *Discrete Contin. Dyn. Syst.-Ser. A*, vol. 38, no. 4, pp. 1719-1744, 2018.

[23] L. Boudin, F. Salvarani, and E. Trélat, "Exponential convergence towards consensus for non-symmetric linear first-order systems in finite and infinite dimensions," *SIAM J. Math. Anal.*, vol. 54, no. 3, pp. 2727-2752, 2022.

[24] B. Bonnet, N. P. Duteil, and M. Sigalotti, "Consensus formation in first-order graphon models with time-varying topologies," arXiv: 2111.03900, 2022.

[25] H. P. McKean, "Propagation of chaos for a class of non-linear parabolic equations," *Lecture Series in Differential Equations*, vol. 7, pp. 41-57, 1967.

[26] A. S. Sznitman, "Topics in propagation of chaos," *Ecole d'été de probabilités de Saint-Flour XIX-1989*, pp. 165-251, 1991.

[27] B. $\emptyset$ ksendal, *Stochastic Differential Equations: An Introduction With Applications*, 6th ed. Berlin: Springer, 2003.

[28] A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*. New York: Academic Press, 1977.