SIAM J. CONTROL OPTIM.
Vol. 53, No. 6, pp. 3339–3363

STOCHASTIC APPROXIMATION BASED CONSENSUS DYNAMICS OVER MARKOVIAN NETWORKS*

MINYI HUANG†, TAO LI‡, AND JI-FENG ZHANG§

Abstract. This paper considers consensus problems with random networks. A key object of our analysis is a sequence of stochastic matrices which involve Markovian switches and decreasing step sizes. We establish ergodicity of the backward products of these stochastic matrices. The basic technique is to consider the second moment dynamics of an associated Markovian jump linear system and exploit its two-scale interaction property resulting from the decreasing step sizes. The mean square convergence rate of the backward products is also obtained. The ergodicity results are used to prove mean square consensus of stochastic approximation algorithms where agents collect noisy information. The approach is further applied to a token scheduled averaging model.

Key words. backward product, consensus, ergodicity, Markovian switch, mean square convergence, stochastic approximation

AMS subject classifications. 93E03, 93E15, 94C15, 68R10, 60J10

DOI. 10.1137/140984348

1. Introduction.

Consensus algorithms with imperfect information exchange or randomly perturbed state evolution have been systematically investigated, addressing many important issues including measurement noise, exogenous perturbations entering system state dynamics, and the quantization effect [1, 12, 13, 15, 24, 30]. The work [28] made an early effort introducing stochastic gradient based consensus algorithms. For noisy modeling of collective motion in multiagent systems, see, e.g., [6, 22].

When noisy measurements of neighboring agents' states are available, stochastic approximation with decreasing step sizes may be applied to reduce long-term fluctuation of the iteration [12, 13, 18, 19, 23, 26]. A popular tool for proving convergence is to use quadratic Lyapunov functions. For fixed network topologies containing a spanning tree, the existence of such functions is guaranteed. This is provable by the constructive method in [11, 31]. For time-varying topologies, the use of Lyapunov functions typically depends on assuming balanced graphs or restrictive eigenvalue

*Received by the editors August 29, 2014; accepted for publication (in revised form) July 20, 2015; published electronically November 12, 2015. A preliminary version of this paper appeared in *Proceedings of the 53rd IEEE Conference on Decision and Control*, Los Angeles, CA, 2014, pp. 2216–2221.

http://www.siam.org/journals/sicon/53-6/98434.html

†School of Mathematics and Statistics, Carleton University, Ottawa, ON K1S 5B6, Canada (mhuang@math.carleton.ca). The work of this author was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under a discovery grant and a discovery accelerator supplements program. Part of his work was performed when he visited the Chinese Academy of Sciences, Beijing, during September 2013–January 2014.

‡School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China (sixumuzi@shu.edu.cn). The work of this author was supported by the National Natural Science Foundation of China under grant 61370030, the Shanghai Rising-Star Program under grant 15QA1402000, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

§Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (jif@iss.ac.cn). The work of this author was supported by the National Natural Science Foundation of China under grant 61120106011 and the National Key Basic Research Program of China (973 Program) under grant 2014CB845301.

conditions [1, 10, 19]. For time-varying directed graphs, the assumption of balanced weights is very restrictive.

To overcome the limitation of the Lyapunov approach, a new technique is introduced in [9]. Consider the stochastic consensus algorithm

$$X_{t+1} = (I + a_t B_t^o)X_t + a_t D_t^o W_t, \quad t \geq 0,$$

for n agents with randomly varying network topologies. Each matrix B_t^o is determined by the random network topology at time t modeled by a directed graph G_t^o. The random matrix sequence $\{D_t^o, t \geq 0\}$ is bounded. The sequence $\{W_t, t \geq 0\}$ consists of independent vector random variables and is independent of $\{(B_t^o, D_t^o), t \geq 0\}$. We take the step size a_t satisfying the standard conditions in stochastic approximation. It is shown that mean square consensus is ensured if and only if $\{A_t := I + a_t B_t^o, t \geq 0\}$ has ergodic backward products with probability one. By studying the trajectory behavior of a switching linear system, it is further shown that such ergodicity holds, and so the balanced graph assumption is removed. This idea is very different from using paracontractions [7] or Wolfowitz's theorem [29]. A key condition used in [9] is that for a zero probability set N_0 and each $\omega \in \Omega \backslash N_0$, there exists a sequence $0 = T_0(\omega) < T_1(\omega) < T_2(\omega) < \dots$ such that the graph union is strongly connected on each discrete time interval $[T_l(\omega), T_{l+1}(\omega))$, $l \geq 0$, and

$$\sup_l [T_{l+1}(\omega) - T_l(\omega)] < \infty. \tag{1.1}$$

In this paper, we are interested in an important class of networks, where the switches are governed by a finite state Markov chain $\{\theta_t, t \geq 0\}$ and condition (1.1) does not hold in general. The Markovian switches can model communication failure [10, 33] and randomized scheduling for signal transmission. Our analysis starts by considering the matrix sequence $\{I + a_t B_{\theta_t}, t \geq 0\}$ which naturally arises in stochastic approximation algorithms in order to attenuate noise and is also of interest in its own right. In some other situations, general stochastic matrices of the form $\{I + a_t B_t^o, t \geq 0\}$, converging to an identity matrix and so referred to as degenerating stochastic matrices [9], can be used to model hardening positions in consensus models [2, 4]. In relation to [9], the route of analyzing stochastic approximation through studying ergodic backward products of $\{I + a_t B_{\theta_t}, t \geq 0\}$ is still valid in this Markovian switching model. However, we need to develop very different techniques to establish ergodicity. We introduce an auxiliary noiseless Markovian jump linear system associated with degenerating stochastic matrices and next examine the dynamics of its second moment matrix, which is similar to [5, 21]. Based on the second moment dynamics, we further identify a class of time-varying linear systems with two-scale interactions, on which we will develop the main machinery for eventually proving ergodicity of backward products. We also obtain the mean square convergence rate of the backward products.

The approach of this paper will be further applied to study a noisy averaging model where a token is used to schedule the broadcast of the state information of a node. A well-known randomized scheduling rule for broadcast is to employ independent Poisson clocks [8, 32]. Our scheduling mechanism has certain advantages since the nodes have more autonomy in their operation. In contrast, Poisson clocks implicitly demand more coordination since all agents should refer to a common time scale.

We mention some recent literature on ergodicity of stochastic matrices over random networks. The work [21] considers backward products of $\{A_{\theta_t}, t \geq 0\}$ and establishes their almost sure convergence by the second moment dynamics. Average consensus is proved when each matrix A_{θ_t} is further assumed to be doubly stochastic [20]. The approach of [21] is to tackle a time-invariant linear difference equation and the main condition is that the Markov chain is irreducible and that the graph union contains a spanning tree. Our model gives rise to a time-varying difference equation for the second moment dynamics and for this reason the associated asymptotic analysis is very different from [21]. For a sequence of independent stochastic matrices $\{A_t, t \geq 0\}$, ergodicity is proved by an infinite flow approach in [27].

The key idea in our two-scale analysis of the second moment dynamics is to construct a lower dimensional model which is able to reflect certain connectivity properties ensured by the graph union. In a different context, two-scale consensus modeling with Markovian regime switching is introduced in [16, 34] and weak convergence analysis is developed. The model in [16] includes a faster Markov chain to be tracked by multiple sensors. The work [34] treats different relative values of the regime switching rate and the step size used in the state update. For multiagent parameter estimation problems, [14] uses step sizes of different scales for averaging states and incorporating local parameter estimation.

We make some notes on notation. We use 1_k to denote a column vector consisting of k ones, and $J_k = \frac{1}{k}1_k1_k^T$. The indicator function of an event A is denoted by 1_A. We use I to denote an identity matrix with its dimension clear from the context. For clarity, we sometimes indicate the dimension by adding a subscript (such as k in I_k). The number $M(i,j)$ denotes the (i,j)th entry of a matrix M. For a vector or matrix M, denote the Frobenius norm $|M| = [\mathrm{Tr}(M^TM)]^{1/2}$. For column vectors $Z_1, \dots, Z_k$, $[Z_1; \dots; Z_k]$ denotes the column vector obtained by vertical concatenation of the k vectors. Let $\{g_t, t \geq 0\}$ and $\{h_t, t \geq 0\}$ be two sequences where the latter is a nonnegative sequence. Then $g_t = O(h_t)$ means that there exist constants C and T such that $|g_t| \leq Ch_t$ for all $t \geq T$, and $g_t = o(h_t)$ means that for any $\epsilon > 0$, there exists T such that $|g_t| \leq \epsilon h_t$ for all $t \geq T$. The agent or node index is often used as a superscript (x_t^i, ζ_t^j, κ_t^i, etc.) and should not be understood as an exponent. We also write some vectors ($\phi_t^k, z_t^k \in \mathbb{R}^N$) with superscript k, which is obviously seen not to be an exponent. The identification of these widely used superscripts should be clear from the context.

The paper is organized as follows. Section 2 introduces the stochastic matrix model with Markovian switches and decreasing step sizes, and section 3 presents the main results on ergodicity and stochastic approximation. Section 4 analyzes the second moment dynamics, and a two-scale model is obtained in section 5. Section 6 develops its convergence analysis. Section 7 analyzes the mean square convergence rate of the backward products. An application of the main result in section 3 is presented in section 8, which deals with a token scheduled averaging model. Section 9 concludes the paper.

2. The Markovian switching model.

2.1. Graph theoretic preliminaries. We introduce some standard preliminaries on graph modeling of the network topology. A directed graph (digraph) $G = (\mathcal{N}, \mathcal{E})$ consists of a set of nodes $\mathcal{N} = \{1, \dots, n\}$ and a set of directed edges $\mathcal{E}$. A directed edge (simply called an edge) is denoted by an ordered pair $(i,j) \in \mathcal{N} \times \mathcal{N}$, where $i \neq j$. A directed path (from node i_1 to node i_l) consists of a sequence of nodes $i_1, \dots, i_l$, $l \geq 2$, such that $(i_k, i_{k+1}) \in \mathcal{E}$. The digraph G is strongly connected if from any node to any other node, there exists a directed path. A directed tree is a digraph where each node i, except the root, has exactly one parent node j so that $(j,i) \in \mathcal{E}$.

We call $G' = (\mathcal{N}', \mathcal{E}')$ a subgraph of G if $\mathcal{N}' \subset \mathcal{N}$ and $\mathcal{E}' \subset \mathcal{E}$. The digraph G is said to contain a spanning tree if there exists a directed tree $G_{\rm tr} = (\mathcal{N}, \mathcal{E}_{\rm tr})$ as a subgraph of G. If $(j,i) \in \mathcal{E}$, j is called an in-neighbor (or neighbor) of i, and i is called an out-neighbor of j. Denote $\mathcal{N}_i = \{j | (j,i) \in \mathcal{E}\}$. If G is an undirected graph, each edge is denoted as an unordered pair (i,j), where $i \neq j$.

For a matrix $M = (m_{ij})_{i,j\le k} \in \mathbb{R}^{k\times k}$, if it either is a stochastic matrix or has zero row sums and nonnegative off-diagonal entries, we define its interaction graph as a digraph denoted by $\text{graph}(M) = (\mathcal{N}_M, \mathcal{E}_M)$, where $\mathcal{N}_M = \{1, \dots, k\}$ and $(j,i) \in \mathcal{E}_M$ if and only if $m_{ij} > 0$.

2.2. The Markovian model. Let the underlying probability space be denoted by $(\Omega, \mathcal{F}, P)$. Suppose that $\{\theta_t, t = 0, 1, 2, \dots\}$ is a Markov chain with state space $\{1, \dots, N\}$ and transition probability matrix

$$P_\theta = (p_{lm})_{1\le l,m\le N}.$$

Let $\{B_k, k = 1, \dots, N\}$ be $n \times n$ matrices. Each B_k has zero row sums and nonnegative off-diagonal entries and can be interpreted as the generator of an n state continuous time Markov chain. Each B_k is associated with its interaction digraph $G_k = (\mathcal{N}, \mathcal{E}_k)$, where $\mathcal{N} = \{1, \dots, n\}$ and $(j,i) \in \mathcal{E}_k$ if and only if $b_{ij} > 0$.

Consider the sequence of matrices

$$\{I + a_t B_{\theta_t}, t \ge 0\}.$$

As $t \to \infty$, $I + a_t B_{\theta_t}$ tends to the identity matrix describing a trivial Markov chain without transitions. Following [9], we call it a sequence of degenerating stochastic matrices. Denote the backward product $\Psi_{t,s} = (I + a_{t-1}B_{\theta_{t-1}}) \cdots (I + a_s B_{\theta_s})$ for $t > s$ and $\Psi_{s,s} = I$. Our first task is to examine the asymptotic property of $\Psi_{t,s}$ for any fixed s when $t \to \infty$.

Remark 1. Throughout the paper we assume

$$\inf_{t\ge 0, l, i} (1 + a_t B_l(i,i)) \ge 0$$

and otherwise may start with a large fixed initial time t_0 instead of time 0 and consider $t \ge t_0$.

We make the following assumptions:

(A1) $\{a_t, t = 0, 1, 2, \dots\}$ is a nonnegative sequence satisfying (i) $\sum_{t=0}^\infty a_t = \infty$, (ii) $\sum_{t=0}^\infty a_t^2 < \infty$.

(A2) The Markov chain $\{\theta_t, t \ge 0\}$ with state space $\{1, \dots, N\}$ is ergodic (i.e., irreducible and aperiodic).

The initial distribution of $\{\theta_t, t \ge 0\}$ is fixed and denoted by μ_{θ_0}. By (A2), the Markov chain has a unique stationary distribution $\pi = (\pi_1, \dots, \pi_N)$ consisting of N positive entries [25].

(A3) The union graph $\cup_{k=1}^N G_k$ contains a spanning tree $G_{\cup,\rm tr}$.

3. Ergodicity and stochastic approximation. A sequence of stochastic matrices $\{A_t, t \ge 0\}$ has ergodic backward products if for any given s, $\lim_{t\to\infty} A_t \dots A_{s+1} A_s$ exists and is a matrix of identical rows.

THEOREM 3.1. *Assume* (A1)–(A3). *The sequence of stochastic matrices* $\{I + a_t B_{\theta_t}, t \ge 0\}$ *has ergodic backward products with probability one.*

Before being able to prove this basic result, we need to develop the analytical tools in sections 4–6. The proof of Theorem 3.1 is postponed to Appendix B.

The ergodicity analysis for $\{I + a_t B_{\theta_t}, t \geq 0\}$ on one hand is important for establishing the mean square consensus result in Theorem 3.4 and on the other hand is interesting in its own right.

3.1. Stochastic approximation. Denote $X_t = [x_t^1, \dots, x_t^n]^T$. Consider the stochastic approximation based consensus algorithm

$$X_{t+1} = (I + a_t B_{\theta_t})X_t + a_t D_{\theta_t} W_t, \qquad t \geq 0, \tag{3.1}$$

where the Markov chain $\{\theta_t, t \geq 0\}$ determines the underlying network topology for information exchange between the agents. The dimension of the constant matrices $\{D_1, \dots, D_N\}$ is compatible with the noise vector W_t. This conceptually simple modeling can characterize the temporal correlation in the evolution of the network.

A similar Markovian switching noisy consensus model has been studied in [10]. However, that work assumed either balanced graphs or, more restrictively, the existence of a common Lyapunov function. The present work does not depend on such assumptions.

(A4) $\{W_t, t \geq 0\}$ is a sequence of independent vector random variables of zero mean, which is independent of $\{\theta_t, t \geq 0\}$. In addition, $\sup_t E|W_t|^2 < \infty$ and $E|X_0|^2 < \infty$.

To study the convergence of (3.1), we introduce the definition.

Definition 3.2. *The n nodes are said to achieve mean square consensus if $E|x_t^i|^2 < \infty$, $t \geq 0$, $1 \leq i \leq n$, and there exists a random variable x^* such that $\lim_{t\to\infty} E|x_t^i - x^*|^2 = 0$ for $1 \leq i \leq n$.*

The next lemma is an immediate consequence of [9, Theorem 3] by running (3.1) with a general initial time-state pair (t_0, X_{t_0}), $t_0 \geq 0$.

Lemma 3.3. *Under* (A1)–(A4), (3.1) *ensures mean square consensus for any given initial time-state pair (t_0, X_{t_0}) with $E|X_{t_0}|^2 < \infty$ if and only if $\{I + a_t B_{\theta_t}\}$ has ergodic backward products with probability one.*

Theorem 3.4. *Assume* (A1)–(A4). *The algorithm* (3.1) *ensures mean square consensus.*

Proof. This theorem follows from Lemma 3.3 and Theorem 3.1. ◻

4. The second moment dynamics. Throughout this section, (A1)–(A2) are assumed. The backward products of $\{I + a_t B_{\theta_t}, t \geq 0\}$ will be studied by use of the difference equation

$$X_{t+1} = (I + a_t B_{\theta_t})X_t. \tag{4.1}$$

For this linear system, we run it with any initial time-state pair (t_0, X_{t_0}), where X_{t_0} is deterministic. The process $\{\theta_t, t \geq t_0\}$ is the restriction of the original Markov chain $\{\theta_t, t \geq 0\}$ on the discrete time interval $[t_0, \infty)$. For $t \geq 0$, let μ_{θ_t} be the distribution of θ_t.

Denote

$$V_l(t) = E\left[X_t X_t^T 1_{\{\theta_t = l\}}\right], \quad t \geq t_0, \tag{4.2}$$

$$V(t) = \sum_{l=1}^N V_l(t). \tag{4.3}$$

The expectation in (4.2) is evaluated using $(X_{t_0}, \mu_{\theta_{t_0}})$, where $\mu_{\theta_{t_0}}$ in turn is determined from μ_{θ_0}. The object $V_l(t)$ was also used in [21] for a Markovian switching linear consensus model $X_{t+1} = A_{\theta_t} X_t$ which does not have a step size a_t as in (4.1). The approach of [21] is to obtain a time-invariant linear system for $\{V_l, 1 \le l \le N\}$ and check its asymptotic property, which is very different from our approach to be developed below.

Recall that $\pi = (\pi_1, \dots, \pi_N)$ is the stationary distribution of $\{\theta_t, t \ge 0\}$. For $t \ge t_0$, we have the second moment dynamics

$$\begin{aligned} V_l(t+1) &= E\left[X_{t+1}X_{t+1}^T 1_{\{\theta_{t+1}=l\}}\right] \\ &= \sum_{m=1}^N E\left[(I + a_t B_m)X_t X_t^T (I + a_t B_m)^T 1_{\{\theta_{t+1}=l,\theta_t=m\}}\right] \\ &= \sum_{m=1}^N p_{ml} E\left[(I + a_t B_m)X_t X_t^T (I + a_t B_m)^T 1_{\{\theta_t=m\}}\right] \\ &= \sum_{m=1}^N p_{ml}(I + a_t B_m)V_m(t)\left(I + a_t B_m^T\right). \end{aligned}$$

For an $m \times n$ matrix M, $\mathrm{vec}(M)$ is an mn dimensional column vector obtained by stacking its n columns in order with the first column on top. Let $\xi_t^l = \mathrm{vec}(V_l(t))$ and $\xi_t = [\xi_t^1; \dots; \xi_t^N]$ as vertical concatenation of the N components. Denote the Kronecker sum $A \oplus B = A \otimes I_n + I_n \otimes B$ for $n \times n$ matrices A and B. We have

$$\begin{aligned} \xi_{t+1} &= \begin{pmatrix} p_{11}I_{n^2} & p_{21}I_{n^2} & \dots & p_{N1}I_{n^2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1N}I_{n^2} & p_{2N}I_{n^2} & \dots & p_{NN}I_{n^2} \end{pmatrix} \xi_t \\ &\quad + a_t \begin{pmatrix} p_{11}(B_1 \oplus B_1) & \dots & p_{N1}(B_N \oplus B_N) \\ \vdots & \ddots & \vdots \\ p_{1N}(B_1 \oplus B_1) & \dots & p_{NN}(B_N \oplus B_N) \end{pmatrix} \xi_t \\ &\quad + a_t^2 \begin{pmatrix} p_{11}(B_1 \otimes B_1) & \dots & p_{N1}(B_N \otimes B_N) \\ \vdots & \ddots & \vdots \\ p_{1N}(B_1 \otimes B_1) & \dots & p_{NN}(B_N \otimes B_N) \end{pmatrix} \xi_t \\ &=: (M_{1,0} + a_t M_{2,0} + a_t^2 M_{3,0})\xi_t. \end{aligned} \tag{4.4}$$

A matrix is said to be nonnegative if all its entries are nonnegative.

Proposition 4.1. (i) *Both $M_{2,0}$ and $M_{3,0}$ have zero row sums.* (ii) *$M_{1,0} + a_t M_{2,0} + a_t^2 M_{3,0}$ is a nonnegative matrix.*

Proof. Part (i) can be verified directly. We check (ii). By Remark 1, the only possible entries within $a_t M_{2,0} + a_t^2 M_{3,0}$ to have negative values are the $((l-1)n^2 + i, (m-1)n^2 + i)$th entries, $l, m = 1, \dots, N$, $i = 1, \dots, n^2$. We take $l = 1, m = 1, i = 1$, and all other cases can be checked similarly. The $(1,1)$th entry of the matrix $M_{1,0} + a_t M_{2,0} + a_t^2 M_{3,0}$ is

$$p_{11}[1 + 2b_1(1,1)a_t + (b_1(1,1))^2 a_t^2] \ge 0.$$

This proves (ii). ∎

To facilitate further analysis, we will modify (4.4) into a new form. Denote the matrix $\Pi = \mathrm{diag}(\pi_1 I_{n^2}, \dots, \pi_N I_{n^2}) \in \mathbb{R}^{Nn^2 \times Nn^2}$ and introduce the linear transformation

$$\bar{\xi}_t = \Pi^{-1}\xi_t, \quad t \ge t_0.$$

Denote

$$M_1 = \Pi^{-1} M_{1,0} \Pi = \begin{pmatrix} \frac{\pi_1 p_{11}}{\pi_1} I_{n^2} & \frac{\pi_2 p_{21}}{\pi_1} I_{n^2} & \dots & \frac{\pi_N p_{N1}}{\pi_1} I_{n^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\pi_1 p_{1N}}{\pi_N} I_{n^2} & \frac{\pi_2 p_{2N}}{\pi_N} I_{n^2} & \dots & \frac{\pi_N p_{NN}}{\pi_N} I_{n^2} \end{pmatrix},$$

$$M_2 = \Pi^{-1} M_{2,0} \Pi = \begin{pmatrix} \frac{\pi_1 p_{11}}{\pi_1} (B_1 \oplus B_1) & \dots & \frac{\pi_N p_{N1}}{\pi_1} (B_N \oplus B_N) \\ \vdots & \ddots & \vdots \\ \frac{\pi_1 p_{1N}}{\pi_N} (B_1 \oplus B_1) & \dots & \frac{\pi_N p_{NN}}{\pi_N} (B_N \oplus B_N) \end{pmatrix},$$

and $M_3 = \Pi^{-1} M_{3,0} \Pi$. Then

$$\bar{\xi}_{t+1} = (M_1 + a_t M_2 + a_t^2 M_3) \bar{\xi}_t. \tag{4.5}$$

Although $\bar{\xi}_t$ has been defined in terms of $\{V_l(t), 1 \le l \le N\}$ for $t \ge t_0$, the linear system (4.5) can be studied in terms of any initial pair $(t_1, \bar{\xi}_{t_1}) \in \mathbb{Z}_+ \times \mathbb{R}^{Nn^2}$ for $t_1 \ge 0$.

PROPOSITION 4.2. (i) *M_2 and M_3 have zero row sums.* (ii) *For each $t \ge 0$, $M_1 + a_t M_2 + a_t^2 M_3$ is a stochastic matrix.* (iii) *$M_1 + a_t M_2$ is a stochastic matrix for all $t \ge t_0^*$ provided that* $\inf_{t \ge t_0^*, l, i} (1 + 2a_t B_l(i,i)) \ge 0$.

Proof. Analogous to the proof of Proposition 4.1, we can show that (i) holds. Furthermore, $M_1 + a_t M_2 + a_t^2 M_3$ is a nonnegative matrix. Now it suffices to show that M_1 has unit row sums. For each l, the stationary distribution $(\pi_1, \dots, \pi_N)$ satisfies $\sum_{k=1}^N \pi_k p_{kl} = \pi_l$. So the lth row sum of M_1 equals 1. Part (ii) follows. We check the $((l-1)n^2 + i, (m-1)n^2 + i)$th entry of $M_1 + a_t M_2$, $l, m = 1, \dots, N$, $i = 1, \dots, n^2$. For instance,

$$[M_1 + a_t M_2](1,1) = p_{11}(1 + 2a_t B_1(1,1)) \ge 0,$$
$$[M_1 + a_t M_2](2,2) = p_{11}(1 + a_t B_1(1,1) + a_t B_1(2,2)) \ge 0$$

for $t \ge t_0^*$. In this manner, the $N^2 n^2$ entries are verified to be nonnegative. All remaining entries of $M_1 + a_t M_2$ are clearly nonnegative. Part (iii) follows. ∎

Since $a_t \to 0$, there exists t_0^* satisfying the condition in Proposition 4.2. We consider the new linear system

$$\zeta_{t+1} = (M_1 + a_t M_2) \zeta_t, \qquad t \ge t_0^*. \tag{4.6}$$

We denote two statements: $S1$ (resp., $S2$)—Algorithm (4.5) (resp., (4.6)) ensures consensus with any given initial pair $(t_1, \bar{\xi}_{t_1})$ (resp., $(t_1, \zeta_{t_1}), t_1 \ge t_0^*$).

LEMMA 4.3. *$S1$ is equivalent to $S2$.*

Proof. For given initial pairs $(t_1, \bar{\xi}_{t_1})$ and (t_1, ζ_{t_1}), we have $\sum_{t=t_1}^\infty a_t^2 |M_3 \bar{\xi}_t| < \infty$ and $\sum_{t=t_1}^\infty a_t^2 |M_3 \zeta_t| < \infty$. Thus one algorithm may be viewed as another subject to small perturbation. The method is similar to the proof of [9, Lemma B.2]. ∎

5. The averaging model with two-scale interactions. Throughout this section, (A1)–(A3) are assumed. We view (4.6) as a consensus problem with Nn^2 agents indexed by $\{1, 2, \dots, Nn^2\}$. To identify the interaction relation of these agents, we introduce a small parameter $\epsilon > 0$ and define the matrix

$$M_\epsilon = M_1 + \epsilon M_2.$$

Denote $\beta = \max_{k,i} |B_k(i,i)| > 0$. For each fixed

$$\epsilon \in (0, (4\beta)^{-1}],$$

M_ϵ is a stochastic matrix and can be associated with a Markov chain $\{\Upsilon_t, t \geq 0\}$ of Nn^2 states $\{1, 2, \ldots, Nn^2\}$. Denote the list

$$\begin{aligned} S_1 &= \{1, n^2+1, \ldots, (N-1)n^2+1\},\\ S_2 &= \{2, n^2+2, \ldots, (N-1)n^2+2\},\\ &\vdots\\ S_{n^2} &= \{n^2, 2n^2, \ldots, Nn^2\}. \end{aligned}$$

This list will be used as a partition of the states of $\{\Upsilon_t, t \geq 0\}$ and later on for classifying the Nn^2 agents of (4.6) into n^2 groups.

Denote the matrix

$$P_\pi = (q_{lm})_{1\leq l,m\leq N} = \left(\frac{\pi_m p_{ml}}{\pi_l}\right)_{1\leq l,m\leq N},$$

which can be verified to be a stochastic matrix.

LEMMA 5.1. *The stochastic matrix* $(q_{lm})_{l,m\leq N}$ *is ergodic and its stationary distribution is* π.

Proof. See Appendix A. □

THEOREM 5.2. *Suppose that* (A3) *holds with* i_0 *being the root of* $G_{\cup,\mathrm{tr}}$. *Then the state* i_0 *of the* Nn^2 *state Markov chain* $\{\Upsilon_t, t \geq 0\}$ *is reachable from any other state with positive probability; equivalently,* $\mathrm{graph}(M_\epsilon)$ *contains a spanning tree* $G_{M_\epsilon,\mathrm{tr}}$ *with* i_0 *being its root.*

Proof. See Appendix A. □

For the N states in S_i, $i \leq n^2$, denote the transition probability

$$p_{lm}^{(i)} = P\left(\Upsilon_{t+1} = (m-1)n^2 + i | \Upsilon_t = (l-1)n^2 + i\right)$$

and $P^{(i)} = (p_{lm}^{(i)})_{l,m\leq N}$. It is straightforward to show

$$P^{(i)} = (q_{lm})_{l,m\leq N} + \epsilon Q^{(i)},$$

which is a substochastic matrix and where $Q^{(i)}$ does not depend on ϵ.

Remark 2. When ϵ becomes very small, the transition probabilities among the states within S_i are mainly determined by the ergodic matrix $(q_{lm})_{l,m\leq N}$. By the structure of M_ϵ, the transition probability from one state in S_i to another in S_j, $i \neq j$ (if nonzero) is on the order of ϵ.

We visualize $S_1 \cup \cdots \cup S_{n^2}$ as a decomposition of the state space of $\{\Upsilon_t, t \geq 0\}$ where strong interactions exist within each set S_i and no strong interactions exist between any S_i and S_j, $i \neq j$. Below we will exploit this structure to transform (4.6) into an equivalent form, which appears to be simpler. This will be done using a_t in place of ϵ.

Recall that ζ_t in (4.6) is viewed as the state vector of Nn^2 agents. Denote $\zeta_t = [\zeta_t^1, \zeta_t^2, \ldots, \zeta_t^{Nn^2}]^T$, where each superscript $j \leq Nn^2$ is used as an agent index. Now we rewrite (4.6) by reordering the position of the Nn^2 agents. The collection $S_1, \ldots, S_{n^2}$

will be used to denote different groups of the Nn^2 agents of (4.6). Let $\phi_t^k \in \mathbb{R}^N$ be the states of the agents with indices in S_k,

$$\phi_t^k = \left[\zeta_t^k, \zeta_t^{n^2+k}, \dots, \zeta_t^{(N-1)n^2+k}\right]^T, \quad 1 \le k \le n^2. \tag{5.1}$$

We take a permutation of the components of ζ_t to get the new vector

$$\phi_t := \left[\phi_t^1; \phi_t^2; \dots; \phi_t^{n^2}\right].$$

In fact, there exists a unique nonsingular matrix Γ such that

$$\phi_t = \Gamma\zeta_t.$$

By (4.6), the new state vector ϕ_t satisfies

$$\phi_{t+1} = \Gamma(M_1 + a_t M_2)\Gamma^{-1}\phi_t =: \hat{M}(a_t)\phi_t. \tag{5.2}$$

It is clear that $\hat{M}(a_t)$ is a stochastic matrix if $M_1 + a_t M_2$ is.

Remark 3. By Proposition 4.2, $\hat{M}(a_t)$ is a stochastic matrix for all large t.

THEOREM 5.3. *$\hat{M}(a_t)$ has the representation*

$$\hat{M}(a_t) = \begin{bmatrix} \hat{M}_{11}(a_t) & a_t\hat{M}_{12} & \cdots & a_t\hat{M}_{1n^2} \\ a_t\hat{M}_{21} & \hat{M}_{22}(a_t) & \cdots & a_t\hat{M}_{2n^2} \\ \vdots & & \ddots & \\ a_t\hat{M}_{n^21} & a_t\hat{M}_{n^22} & \cdots & \hat{M}_{n^2n^2}(a_t) \end{bmatrix}, \tag{5.3}$$

where

(i) $\hat{M}_{ij} \in \mathbb{R}^{N\times N}$ *is a constant nonnegative matrix for any $i \ne j$, and so independent of the value of a_t.*

(ii) $\hat{M}_{ii}(a_t) + a_t\sum_{j\ne i}\hat{M}_{ij} = (q_{lm})_{l,m\le N}$ *for all $i \le n^2$.*

Proof. Consider (4.6) and any agent $i' \in S_i$ with state of the form $\zeta_t^{(j-1)n^2+i}$ for some $1 \le j \le N$. If this agent updates its state using the state of an agent $j' \in S_j$, the weight assigned to j' can only originate as an entry of $a_t M_2$; see Remark 2. This implies that all off-diagonal blocks in (5.3) must take the form $a_t\hat{M}_{ij}$. By Proposition 4.2, whenever a_t is sufficiently small, $M_1 + a_t M_2$ and so $\hat{M}(a_t)$ are nonnegative matrices. So $\hat{M}_{ij}$ is nonnegative for $i \ne j$. This proves (i).

To show (ii), we check $A_1 := \hat{M}_{11}(a_t) + a_t\sum_{j\ne 1}\hat{M}_{1j}$. First, $\hat{M}_{11}(a_t)(1,1) = q_{11} + 2a_t q_{11} B_1(1,1)$. Next, $\sum_{j=2}^{n^2} a_t\hat{M}_{1j}(1,1) = 2a_t q_{11}\sum_{j=2}^{n} B_1(1,j)$. Therefore, $A_1(1,1) = q_{11}$. We continue to check $A_1(1,l)$, $1 < l \le N$. Then

$$\begin{aligned} \hat{M}_{11}(1,l) &= M_1(1,(l-1)n^2+1) + a_t M_2(1,(l-1)n^2+1) \\ &= q_{1l} + a_t q_{1l}(2B_l(1,1)), \\ a_t\hat{M}_{1j}(1,l) &= a_t M_2(1,(l-1)n^2+j) \\ &= a_t q_{1l}(B_l \otimes I_n + I_n \otimes B_l)(1,j), \quad j \ge 2, \end{aligned}$$

which is the weight agent 1 in (4.6) assigns to the agent as the lth member of the jth group S_j. It can be checked that

$$\sum_{j=2}^{n^2}(B_l \otimes I_n + I_n \otimes B_l)(1,j) = 2\sum_{k=2}^{n} B_l(1,k).$$

It follows that

$$A_1(1,l) = \hat{M}_{11}(1,l) + \sum_{j=2}^{n^2} a_t \hat{M}_{1j}(1,l) = q_{1l}.$$

In the same manner we can check the remaining entries of A_1 and also other cases of $\hat{M}_{ii}$, $2 \le i \le n^2 - 1$. The theorem follows. ∎

By Theorem 5.3, we may write (5.2) as

$$\begin{bmatrix} \phi_{t+1}^1 \\ \phi_{t+1}^2 \\ \vdots \\ \phi_{t+1}^{n^2} \end{bmatrix} = \begin{bmatrix} \hat{M}_{11}(a_t) & a_t\hat{M}_{12} & \cdots & a_t\hat{M}_{1n^2} \\ a_t\hat{M}_{21} & \hat{M}_{22}(a_t) & \cdots & a_t\hat{M}_{2n^2} \\ \vdots & & \ddots & \\ a_t\hat{M}_{n^21} & a_t\hat{M}_{n^22} & \cdots & \hat{M}_{n^2n^2}(a_t) \end{bmatrix} \begin{bmatrix} \phi_t^1 \\ \phi_t^2 \\ \vdots \\ \phi_t^{n^2} \end{bmatrix}, \tag{5.4}$$

which will be called a canonical form of (4.6).

We may view (5.4) as a two-scale averaging model. To avoid confusion, when a consensus model is examined with a corresponding number of agents, the index of an agent is specified according to the position of its state within the state vector. For instance, ϕ_t^1 denotes the states of agents with indices $\{1, \ldots, N\}$. Denote $\hat{S}_k = \{(k-1)N+1, \ldots, kN\}$, $k = 1, \ldots, n^2$. By (5.1), it is evident that the agent indices $\hat{S}_k$ in (5.4) and S_k in (4.6) refer to the same group of agents physically.

The canonical form makes it convenient to identify the interaction structure of the Nn^2 agents. Within each group $\hat{S}_k$, averaging takes place rapidly when (5.4) is iterated. The interconnection between the groups is controlled by the step size a_t. By Theorem 5.3, once a_t is fixed, the matrix $\hat{M}(a_t)$ is completely determined by the set of off-diagonal blocks. We will continue to check whether they will be able to generate adequate interactions among the groups $\{\hat{S}_1, \ldots, \hat{S}_{n^2}\}$ in some sense.

We define a new graph which has fewer nodes than graph$(\hat{M}(\epsilon))$. Its purpose is to indicate the information flow among different agent groups $\hat{S}_1, \ldots, \hat{S}_{n^2}$ of (5.4).

Let $\hat{G}_{\mathrm{q}}$ be a digraph with nodes $\mathcal{N}_{\mathrm{q}} = \{1, 2, \ldots, n^2\}$ and the set of edges $\mathcal{E}_{\mathrm{q}}$. An edge $(j,i) \in \mathcal{E}_{\mathrm{q}}$ if and only if $\hat{M}_{ij} \neq 0$. If we identify all nodes of each S_i as an equivalent class, $\hat{G}_{\mathrm{q}}$ defined above may be called a quotient graph of graph(M_ϵ). The graph $\hat{G}_{\mathrm{q}}$ does not depend on the particular value of the small parameter ϵ.

LEMMA 5.4. For $\hat{G}_{\mathrm{q}}$, $(j,i) \in \mathcal{E}_{\mathrm{q}}$ if and only if there is an edge on graph(M_ϵ) from a node in S_j to a node in S_i.

Proof. There is an edge on graph(M_ϵ) from a node in S_j to a node in S_i if and only if $\hat{M}_{ij} \neq 0$. ∎

THEOREM 5.5. *$\hat{G}_{\mathrm{q}}$ contains a spanning tree.*

Proof. By Theorem 5.2, graph(M_ϵ) contains a spanning tree $G_{M_\epsilon,\mathrm{tr}}$. Without loss of generality, assume that the root of $G_{M_\epsilon,\mathrm{tr}}$ is node 1. It suffices to show that node 1 of $\hat{G}_{\mathrm{q}}$ can reach any other node $j \in \{2, \ldots, n^2\}$ by a directed path. Select such a node j.

Consider graph(M_ϵ). There exists a directed path from node $1 \in S_1$ to node $j \in S_j$. Denote this directed path by $1, k_2, k_3, \ldots, k_r, j$. Suppose that $k_i \in S_{d_i}$. We list $S_1, S_{d_2}, \ldots, S_{d_r}, S_j$. For this list, if S_k appears successively in a segment, we list S_k only once corresponding to that segment. By Lemma 5.4, the resulting list identifies a directed path from node 1 to node j in $\hat{G}_{\mathrm{q}}$. ∎

Remark 4. Theorems 5.2, 5.3, 5.5, and Lemma 5.4 still hold if (A2) is replaced by the weaker assumption that $\{\theta_t, t \ge 0\}$ is irreducible while all other assumptions remain the same.

6. Convergence of algorithm (5.4).

Assume (A1)–(A3) for this section. For each ϕ_t^k, denote $\phi_t^k = [\phi_t^{k,1}, \ldots, \phi_t^{k,N}]^T \in \mathbb{R}^N$. In this section the integer $k \le n^2$ will frequently be used as a superscript but not an exponent for various vectors. Consider (5.4) with any given initial pair (t_1, ϕ_{t_1}). Our method is to derive a lower dimensional model. Each component ϕ_t^k corresponds to N equations within (5.4) for which we attempt to only retain the equation for $\phi_t^{k,1}$.

Recall that $P_\pi = (q_{lm})_{l,m\le N}$ is an ergodic stochastic matrix. Denote its N eigenvalues by $\lambda_1 = 1, \lambda_2, \ldots, \lambda_N$. Then $\max_{2\le l\le N} |\lambda_l| < 1$. Fix any $\delta \in (\max_{2\le l\le N} |\lambda_l|, 1)$. Define

$$a_t^* = \sum_{s=0}^{t} \delta^{t-s} a_s, \quad t \ge 0.$$

The next lemma provides some prior estimate of the difference between different entries in ϕ_t^k.

Lemma 6.1. *We have*

$$\max_k \max_{l,m} |\phi_t^{k,l} - \phi_t^{k,m}| = O(a_t^*), \quad t \ge t_1.$$

Proof. First, there exists a constant C, depending on the initial pair (t_1, ϕ_{t_1}) of (5.4), such that $\sup_{t,k} |\phi_t^k| \le C$; see Remark 3. Denote $H_k(a_t) = a_t \sum_{j\ne k} \hat{M}_{kj}(\phi_t^j - \phi_t^k)$. Hence $|H_k(a_t)| = O(a_t)$. Next, we check ϕ_t^k and by Theorem 5.3 have the relation

$$\phi_{t+1}^k = P_\pi \phi_t^k + H_k(a_t). \tag{6.1}$$

Note that $P_\pi - I$ has rank $N-1$. Let Φ_{N-1} be an $n \times (n-1)$ matrix such that $\mathrm{span}(\Phi_{N-1}) = \mathrm{span}(P_\pi - I)$. Denote $\Phi = [1_N, \Phi_{N-1}] \in \mathbb{R}^{N\times N}$. By the method in [12] we can show that Φ is nonsingular and

$$\Phi^{-1} P_\pi \Phi = \begin{bmatrix} 1 & 0 \\ 0 & A_\pi \end{bmatrix},$$

where A_π is an $(N-1)\times(N-1)$ matrix having all eigenvalues with absolute value less than δ. In fact the first row of Φ^{-1} is equal to π. There exists a constant C such that the power of A_π satisfies

$$|A_\pi^t| \le C\delta^t, \quad t \ge 0.$$

Take a change of coordinates $z_t^k = \Phi^{-1}\phi_t^k \in \mathbb{R}^N$, and denote $z_t^k = [z_t^{k,1}, \ldots, z_t^{k,N}]^T = [z_t^{k,1}; z_t^{k,-1}]$. Thus, $z_t^{k,1} = \pi\phi_t^k$. We obtain

$$\begin{aligned} z_{t+1}^{k,1} &= z_t^{k,1} + O(a_t), \\ z_{t+1}^{k,-1} &= A_\pi z_t^{k,-1} + H_{k,-1}(a_t), \end{aligned} \tag{6.2}$$

where $H_{k,-1}(a_t)$ is determined from $H_k(a_t)$ and so $|H_{k,-1}(a_t)| = O(a_t)$. The second equation leads to

$$\begin{aligned} |z_t^{k,-1}| &= \left| A_\pi^{t-t_1} z_{t_1}^{k,-1} + \sum_{s=t_1}^{t-1} A_\pi^{t-1-s} H_{k,-1}(a_s) \right| \\ &= O\left(\delta^{t-t_1} + a_{t-1}^*\right) = O(a_t^*). \end{aligned} \tag{6.3}$$

Now for $t \geq t_1$,

$$\phi_t^k = \Phi z_t^k = [1_N, \Phi_{N-1}] z_t = z_t^{k,1} 1_N + \Phi_{N-1} z_t^{k,-1}. \tag{6.4}$$

The lemma follows. ∎

For a matrix M, we use $\mathrm{rsum}_l(M)$ to denote the sum of its lth row. With a slight abuse of notation, we will sometimes use $O(a_t)$ (or $o(a_t)$, $O(a_t^*)$, etc.) to denote a vector or matrix of compatible dimension. It means that each entry of the vector or matrix is of the form $O(a_t)$ (or $o(a_t)$, $O(a_t^*)$).

THEOREM 6.2. *For $k = 1, 2, \ldots, n^2$, we have*

$$z_{t+1}^{k,1} = \left(1 + a_t \hat{b}_{kk}\right) z_t^{k,1} + a_t \sum_{j=1, j\neq k}^{n^2} \hat{b}_{kj} z_t^{j,1} + O((a_t^*)^2), \quad t \geq t_1, \tag{6.5}$$

where $\hat{b}_{kj} = \sum_{l=1}^N \pi_l \mathrm{rsum}_l(\hat{M}_{kj})$ for $j \neq k$, and $\hat{b}_{kk} = -\sum_{j=1, j\neq k}^{n^2} \hat{b}_{kj}$.

Proof. By (6.1), we have

$$\begin{aligned} z_{t+1}^{k,1} &= \pi \phi_{t+1}^k \\ &= \pi P_\pi \phi_t^k - a_t \pi \sum_{j\neq k} \hat{M}_{kj} \phi_t^k + a_t \pi \sum_{j \neq k} \hat{M}_{kj} \phi_t^j \\ &= z_t^{k,1} - a_t \pi \sum_{j\neq k} \hat{M}_{kj} \phi_t^k + a_t \pi \sum_{j\neq k} \hat{M}_{kj} \phi_t^j. \end{aligned} \tag{6.6}$$

Since $z_t^{k,1} = \pi \phi_t^k$, it follows from Lemma 6.1 that for each l,

$$|\phi_t^{k,l} - z_t^{k,1}| = \left| \sum_{m=1}^N \pi_m (\phi_t^{k,l} - \phi_t^{k,m}) \right| = O(a_t^*).$$

Therefore,

$$\begin{aligned} \pi \sum_{j\neq k} \hat{M}_{kj} \phi_t^k &= \pi \sum_{j\neq k} \hat{M}_{kj} \left[z_t^{k,1} 1_N + O(a_t^*) \right] = z_t^{k,1} \sum_{j\neq k} \sum_{l=1}^N \pi_l \mathrm{rsum}_l(\hat{M}_{kj}) + O(a_t^*) \\ &= \Big(\sum_{j\neq k} \hat{b}_{kj} \Big) z_t^{k,1} + O(a_t^*). \end{aligned}$$

Similarly,

$$\begin{aligned} \pi \sum_{j\neq k} \hat{M}_{kj} \phi_t^j &= \pi \sum_{j\neq k} \hat{M}_{kj} \left[z_t^{j,1} 1_N + O(a_t^*) \right] = \sum_{j\neq k} z_t^{j,1} \sum_{l=1}^N \pi_l \mathrm{rsum}_l(\hat{M}_{kj}) + O(a_t^*) \\ &= \sum_{j\neq k} \hat{b}_{kj} z_t^{j,1} + O(a_t^*). \end{aligned}$$

The theorem follows by combining (6.6) with the above estimates and the fact that $a_t = O(a_t^*)$. ∎

Remark 5. Lemma 6.1 and Theorem 6.2 hold under a much weaker condition on $a_t, t \geq 0$. We only need $0 \leq a_t \to 0$ and $\sup_t a_t > 0$; the new condition does not affect Theorem 5.3 and it ensures that (6.3) holds and that $M_1 + a_t M_2$ is a stochastic matrix for all large t.

Define

$$\hat{B} = \left(\hat{b}_{kj}\right)_{k,j\le n^2}, \tag{6.7}$$

which has zero row sums and nonnegative off-diagonal entries. Denote $y_t = [z_t^{1,1}, \dots, z_t^{n^2,1}]^T$. Let (6.5) be written in the vector form

$$y_{t+1} = \left(I_{n^2} + a_t\hat{B}\right) y_t + O\left((a_t^*)^2\right), \quad t \ge t_1. \tag{6.8}$$

LEMMA 6.3. graph$(\hat{B}) = \hat{G}_{\mathrm{q}}$.

Proof. Both digraphs have the set of nodes $\{1, \dots, n^2\}$. Note that $\hat{M}_{kj}$ is a nonnegative matrix for any (j,k). Also, the stationary distribution π has N positive entries. So $\hat{b}_{kj} > 0$ if and only if $\hat{M}_{kj} \neq 0$. On the other hand, (j,k) is an edge of graph$(\hat{B})$ if and only if $\hat{b}_{kj} > 0$; (j,k) is an edge of $\hat{G}_{\mathrm{q}}$ if and only if $\hat{M}_{kj} \neq 0$. We conclude that both graph$(\hat{B})$ and $\hat{G}_{\mathrm{q}}$ have the same set of edges. ∎

THEOREM 6.4. *The algorithm* (5.4) *ensures consensus for any give initial pair* (t_1, ϕ_{t_1}).

Proof. Consider the algorithm

$$y'_{t+1} = \left(I_{n^2} + a_t\hat{B}\right) y'_t. \tag{6.9}$$

This is a special case of the stochastic approximation algorithm in [12] by setting the noise as zero. By Theorem 5.5, Lemma 6.3, and the step size condition (A1), (6.9) ensures consensus with any initial pair (t_0, y'_{t_0}).

Given any initial pair (t_1, ϕ_{t_1}), we accordingly determine the initial pair (t_1, y_{t_1}) in (6.8). Denote $r_t = \frac{1-\delta^{t+1}}{1-\delta}$. We observe that

$$(a_t^*)^2 = r_t^2 \left(\sum_{s=0}^{t} \frac{\delta^{t-s}}{r_t} a_s\right)^2 \le r_t^2 \sum_{s=0}^{t} \frac{\delta^{t-s}}{r_t} a_s^2 \le \frac{1}{1-\delta}\sum_{s=0}^{t} \delta^{t-s} a_s^2.$$

This implies that

$$\sum_{t=0}^{\infty} (a_t^*)^2 \le \frac{1}{1-\delta}\left(\sum_{k=0}^{\infty} \delta^k\right)\sum_{s=0}^{\infty} a_s^2 = \frac{1}{(1-\delta)^2}\sum_{s=0}^{\infty} a_s^2 < \infty. \tag{6.10}$$

By the convergence of (6.9), it follows from (6.10) and [9, Lemmas B.1, B.2] that for (6.8) with any given initial pair (t_1, y_{t_1}), y_t converges to a limit vector in span$\{1_{n^2}\}$. In other words, there exists a common constant c such that

$$\lim_{t\to\infty} z_t^{k,1} = c \qquad \text{for all } k = 1, \dots, n^2.$$

Subsequently, $\lim_{t\to\infty} \phi_t^k = \lim_{t\to\infty} \Phi z_t^k = c1_N$ since $\lim_{t\to\infty} |z_t^{k,-1}| = 0$. This gives $\lim_{t\to\infty} \phi_t = c1_{Nn^2}$. The theorem follows. ∎

7. Convergence rate.

Ergodicity of the backward products of $\{I + a_t B_{\theta_t}, t \ge 0\}$ has a central role in analyzing the stochastic approximation algorithm (3.1). Theorem 3.1 only characterizes a qualitative property of the sequence of backward products. Here we aim to obtain more information on its asymptotic behavior by establishing its mean square convergence rate.

With some regularity on $\{a_t, t \geq 0\}$, we may simplify the estimates in section 6. We prove the lemma below without requiring (A1).

LEMMA 7.1. *If* $\{a_t, t \geq 0\}$ *satisfies* $0 < a_t \to 0$ *and* $\lim_{t\to\infty} \frac{a_t}{a_{t+1}} = 1$, *then for any initial pair* (t_1, y_{t_1}),

$$y_{t+1} = \left(I_{n^2} + a_t \hat{B}\right) y_t + O(a_t^2), \quad t \geq t_1, \tag{7.1}$$

where $y_t = [z_t^{1,1}, \dots, z_t^{n^2,1}]^T$ *and* $\hat{B}$ *is defined by* (6.7).

Proof. We follow the notation in section 6 and recall Remark 5. Rewrite (6.2) in the form

$$a_{t+1}^{-1} z_{t+1}^{k,-1} = A_\pi \left(a_t^{-1} z_t^{k,-1}\right) \frac{a_t}{a_{t+1}} + a_{t+1}^{-1} H_{k,-1}(a_t), \quad t \geq t_1.$$

Denote $v_t = a_t^{-1} z_t^{k,-1}$. This gives

$$v_{t+1} = (1 + o(1)) A_\pi v_t + O(1). \tag{7.2}$$

Since A_π is stable (i.e., all its eigenvalues are inside the unit circle), we may specify any $Q_0 > 0$ and solve a unique $P_0 > 0$ from the Lyapunov equation $A_\pi^T P_0 A_\pi - P_0 + Q_0 = 0$. By use of (7.2), we may find a small constant $0 < c_0 < 1$ such that

$$v_{t+1}^T P_0 v_{t+1} \leq (1 - c_0) v_t^T P_0 v_t + O(1).$$

Hence $\sup_{t \geq t_1} |v_t| < \infty$ and

$$|z_t^{k,-1}| = O(a_t). \tag{7.3}$$

By adapting the proofs of Lemma 6.1 and Theorem 6.2, we see that under the current assumption, Lemma 6.1 and consequently (6.5) still hold when a_t^* is replaced by a_t. This completes the proof. ∎

Taking $\gamma \in (1/2, 1]$, we choose

$$a_t = \frac{1}{t^\gamma}, \qquad t \geq 1, \tag{7.4}$$

and $a_0 > 0$. The more general case of $a_t = \frac{c}{t^\gamma}$ for $t \geq 1$, $c > 0$ can be reduced to (7.4) by replacing $\{B_1, \dots, B_N\}$ by a new set of matrices. It is clear that (7.4) satisfies the assumption on $\{a_t, t \geq 0\}$ in Lemma 7.1.

Denote the backward product

$$\Psi_{t+1,t_0} = (I + a_t B_{\theta_t}) \dots (I + a_{t_0} B_{\theta_{t_0}}), \quad t \geq t_0,$$

$\Psi_{t_0,t_0} = I$, where $\{a_t, t \geq 0\}$ is given by (7.4). According to Remark 1, we still assume that $I + a_t B_{\theta_t}$ is a stochastic matrix for all t. Under (A1)–(A3), Theorem 3.1 shows that Ψ_{t+1,t_0} converges with probability one to a random matrix denoted by Ψ_{∞,t_0} which has identical rows. Since graph$(\hat{B})$ contains a spanning tree by Theorem 5.5 and Lemma 6.3, $\hat{B}$ has 1 eigenvalue equal to zero and $n^2 - 1$ eigenvalues having strictly negative real parts [12]. Suppose that $\sigma_0 > 0$ is a constant such that all nonzero eigenvalues of $\hat{B}$ have a real part strictly less than $-\sigma_0$.

THEOREM 7.2. *Let the step sizes be given by* (7.4) *and assume* (A2)–(A3).

(i) *If* $1/2 < \gamma < 1$, *we have*

$$E|\Psi_{t+1,t_0} - \Psi_{\infty,t_0}|^2 = O\left(\frac{1}{t^{2\gamma-1}}\right).$$

(ii) *If* $\gamma = 1$,

$$E|\Psi_{t+1,t_0} - \Psi_{\infty,t_0}|^2 = O\left(\frac{1}{t^{\eta}}\right).$$

where $\eta = \min\{1, \sigma_0\}$.

Proof. Step 1. Consider the linear system

$$X_{t+1} = (I + a_t B_{\theta_t})X_t, \quad t \geq t_0.$$

As in (B.2), set the initial condition $X_{t_0}^{(i)} = e_i$ and denote the corresponding solution $X_t^{(i)} = \Psi_{t,t_0} X_{t_0}^{(i)}$ for $t \geq t_0$. Then

$$\Psi_{t+1,t_0} = \left[X_{t+1}^{(1)}, \ldots, X_{t+1}^{(n)}\right], \quad t \geq t_0.$$

It follows that with probability one $X_t^{(i)}$ converges to $\eta_i 1_n$ which is equal to the ith column of Ψ_{∞,t_0} and where η_i is a random variable. We have

$$|\Psi_{t+1,t_0} - \Psi_{\infty,t_0}|^2 = \sum_{i=1}^{n} \left|X_{t+1}^{(i)} - \eta_i 1_n\right|^2.$$

Below we check $X_t^{(1)}$ and simply write it as $X_t = [X_{t,1}, \ldots, X_{t,n}]^T$. Since $\eta_1 1_n$ is obtained as the limit state vector of a consensus model, we necessarily have

$$\min_k X_{t,k} \leq \eta_1 \leq \max_k X_{t,k}, \quad t \geq t_0.$$

Consequently, $|X_{t,k} - \eta_1| \leq \max_j |X_{t,k} - X_{t,j}| \leq \sum_{j=1}^n |X_{t,k} - X_{t,j}|$ almost surely. We need to estimate $E|X_{t,k} - X_{t,j}|^2$. For the initial condition $X_{t_0} = e_1$, we accordingly define $V_l(t)$ by (4.2) and $\bar{\xi}_t$ by (4.5) for $t \geq t_0$. The cases of $X^{(i)}$, $i \geq 2$, can be handled in exactly the same manner.

Step 2. Recalling (7.1), we write

$$y_{t+1} = \left(I_{n^2} + a_t \hat{B}\right) y_t + O(a_t^2),$$

for which we set the initial time t_0. By an appropriate change of coordinates $y_t = \hat{\Phi} p_t$ [12], we have

$$\begin{aligned}
p_{t+1}^{(1)} &= p_t^{(1)} + O(a_t^2), \\
p_{t+1}^{(-1)} &= \left(I_{n^2-1} + a_t \hat{B}_0\right) p_t^{(-1)} + O(a_t^2),
\end{aligned}$$

where $p_t = [p_t^{(1)}; p_t^{(-1)}]$, $p_t^{(-1)} \in \mathbb{R}^{n^2-1}$ and $\hat{B}_0$ is an $(n^2-1) \times (n^2-1)$ Hurwitz matrix. We have the limits $p_t^{(1)} \to p_\infty^{(1)}$ and $p_t^{(-1)} \to 0$ as $t \to \infty$. For $\{a_t, t \geq 0\}$ given by (7.4), denote $\epsilon_t = \sum_{s=t}^{\infty} a_s^2$. Then

$$\left|p_t^{(1)} - p_\infty^{(1)}\right| = O(\epsilon_t) = O\left(t^{1-2\gamma}\right).$$

Denote $\delta_t = |p_t^{(-1)}|$. There exists a constant c such that $\lim_{t\to\infty} y_t = c1_{n^2}$, and

$$|y_t - c1_{n^2}| = O(\epsilon_t + \delta_t).$$

In other words,

$$\left| \left[z_t^{1,1}, \ldots, z_t^{n^2,1} \right]^T - c1_{n^2} \right| = O(\epsilon_t + \delta_t).$$

By (6.4) and (7.3),

$$|\phi_t^k - c1_N| = O(a_t + \epsilon_t + \delta_t).$$

Thus,

$$|\zeta_t - c1_{Nn^2}| = O(a_t + \epsilon_t + \delta_t).$$

The above estimate is valid for any given (t_0, ζ_{t_0}) and it allows us to have $t_0 \le t_0^*$ in (4.6).

Step 3. For X_t in Step 1 with initial pair (t_0, e_1), we determine $V_l(t_0)$ and accordingly $\bar{\xi}_{t_0}$ for (4.5). Denote the limit of $\bar{\xi}_t$ by $c_1 1_{Nn^2}$ which exists. By setting $\zeta_{t_0} = \bar{\xi}_{t_0}$ in (4.5)–(4.6) and comparing the two solutions, we further obtain

$$|\bar{\xi}_t - c_1 1_{Nn^2}| = O(a_t + \epsilon_t + \delta_t).$$

Let Π be defined as in section 4. It follows that

$$|\xi_t - c_1 \Pi 1_{Nn^2}| = |\Pi \bar{\xi}_t - c_1 \Pi 1_{Nn^2}| = O(a_t + \epsilon_t + \delta_t).$$

On the other hand,

$$\left| V_l(t) - c_1 \pi_l 1_n 1_n^T \right| = \left| \xi_t^l - c_1 \pi_l 1_{n^2} \right| \le |\xi_t - c_1 \Pi 1_{Nn^2}|.$$

Let $V(t)$ be defined by (4.3) and recall $J_n = \frac{1}{n} 1_n 1_n^T$. It follows that

$$|V(t) - c_1 n J_n| = O(a_t + \epsilon_t + \delta_t).$$

Therefore,

$$\begin{aligned} E|X_{t,i} - X_{t,j}|^2 &= (e_i - e_j)^T [V(t) - c_1 n J_n](e_i - e_j) \\ &\le (e_i - e_j)^T |V(t) - c_1 n J_n| (e_i - e_j) \\ &= O(a_t + \epsilon_t + \delta_t). \end{aligned}$$

Step 4. If $1/2 < \gamma < 1$, $\delta_t = O(t^{-\gamma})$ by Lemma A.2. Hence

$$E|X_{t,i} - X_{t,j}|^2 = O(t^{-\gamma} + t^{1-2\gamma}) = O(t^{1-2\gamma}).$$

If $\gamma = 1$, $\delta_t = O(t^{-\eta})$ by Lemma A.2. This gives

$$E|X_{t,i} - X_{t,j}|^2 = O(t^{-1} + t^{-\eta}) = O(t^{-\eta}).$$

By Step 1, the theorem follows. ∎

8. Application to token scheduled averaging. Let $G = (\mathcal{N}, \mathcal{E})$ be a strongly connected digraph, where $\mathcal{N} = \{1, \ldots, n\}$. The token process $\{T_t, t = 0, 1, \ldots\}$ is a random walk on G, and so is a Markov chain with state space $\{1, \ldots, n\}$. Let $\hat{\mu}_0$ be the distribution of T_0. The transition probability is $P(T_{t+1} = j | T_t = i) = \hat{p}_{ij}$, where $\hat{p}_{ij} > 0$ if and only if $i \in \mathcal{N}_j$. Denote $P_T = (\hat{p}_{ij})_{i,j \le n}$. It is evident that G is strongly connected if and only if $\{T_t, t \ge 0\}$ is irreducible.

Each node has a counter κ_t^i, $i \in \mathcal{N}$, $t \ge 0$. The initial value $\kappa_0^i \ge 0$ is a deterministic integer. The counter is updated by the rule

$$\kappa_{t+1}^i = \kappa_t^i + 1_{\{T_{t+1} = i\}}, \quad t \ge 0,$$

where 1_A stands for the indictor function of an event A. This means that the counter is incremented by one upon each new possession of the token.

If $T_t = i$, node i broadcasts its state x_t^i which is received with additive noise by its out-neighbors. If $i \in \mathcal{N}_j$, node j receives the measurement

$$y_t^{ji} = x_t^i + w_t^{ji}, \quad t \ge 0.$$

For convenience of modeling, we define w_t^{ji} for all $(i, j) \in \mathcal{E}$. At time t if no measurement occurs along the edge (i, j), w_t^{ji} is simply included as a dummy random variable. Let $\{a_t, t \ge 0\}$ be a nonnegative step size sequence. When $T_t = i$, the state of node j evolves by the rule

$$x_{t+1}^j = \begin{cases} (1 - a_{\kappa_t^i}) x_t^j + a_{\kappa_t^i} y_t^{ji}, & i \in \mathcal{N}_j, \\ x_t^j, & i \notin \mathcal{N}_j, \end{cases} \quad t \ge 0. \tag{8.1}$$

The above modeling uses t to mark the transitions of the token. There is no need for the nodes to share slotted time. When a node is during a period neither possessing the token nor collecting measurements, it remains in an idle status. Neither its counter nor its state is changed.

For each $i \in \mathcal{N}$, define the matrix $B_i = (B_i(j,k))_{j,k \le n}$ by the following rule. If $i \notin \mathcal{N}_j$, then $B_i(j,k) = 0$ for all k. If $i \in \mathcal{N}_j$,

$$B_i(j,k) = \begin{cases} -1, & k = j, \\ 1, & k = i, \\ 0, & \text{all other } k. \end{cases}$$

For a given $t \ge 0$, we list all random variables $\{w_t^{ji}, (i,j) \in \mathcal{E}\}$ into a vector W_t. The position of w_t^{ji} within W_t is determined only by (i,j). Denote $X_t = [x_t^1, \ldots, x_t^n]^T$. Define $\mathbf{a}_{\kappa_t} = \mathrm{diag}(a_{\kappa_t^1}, \ldots, a_{\kappa_t^n})$. We write (8.1) in the vector form

$$X_{t+1} = (I + \mathbf{a}_{\kappa_t} B_{T_t}) X_t + \mathbf{a}_{\kappa_t} D_{T_t} W_t, \quad t \ge 0, \tag{8.2}$$

where the collection of matrices $\{D_1, \ldots, D_n\}$ can be defined accordingly and we omit the details.

We take $\gamma \in (1/2, 1]$ and

$$a_t = \frac{1}{t^\gamma}, \quad t \ge t_a,$$

for some $t_a \ge 1$ and $a_t \in [0, 1]$ for $t < t_a$. Then for each t, $I + \mathbf{a}_{\kappa_t} B_{T_t}$ is a stochastic matrix.

We introduce the following assumptions for the rest of this section.

(H1) $\{T_t, t \geq 0\}$ is ergodic with stationary distribution $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$.

(H2) $\{W_t, t \geq 0\}$ is a sequence of independent vector random variables of zero mean and $\sup_t E|W_t|^2 < \infty$.

(H3) $\{T_t, t \geq 0\}$ and $\{W_t, t \geq 0\}$ are independent, and $E|X_0|^2 < \infty$.

LEMMA 8.1. *Under* (H1)*, there exists a deterministic constant C such that for each i,*

$$\limsup_{t\to\infty} \frac{|\kappa_t^i - \hat{\pi}_i t|}{\sqrt{t \log\log t}} \leq C.$$

Proof. Consider a fixed i. We write $\kappa_t^i = \kappa_0^i + \sum_{s=1}^t 1_{\{T_s = i\}}$, $t \geq 1$. Following [3, section I.14], let

$$\tau_1 < \tau_2 < \cdots < \tau_k < \cdots$$

be an increasing sequence of all values of $t \geq 1$ for which $T_t = i$. Denote $\rho_k = \tau_{k+1} - \tau_k$, which is called the kth return time. The random variables $\{\rho_k, k \geq 1\}$ are independent and identically distributed [3]. Since $\{T_t, t \geq 0\}$ has finite states and is ergodic,

$$P(\rho_k > s) = O(e^{-\alpha s}) \tag{8.3}$$

for some $\alpha > 0$. Therefore, the finite moment assumptions in [3, Theorem 5, p. 101] hold and the proving argument by the dissection formula implies that there exists C such that

$$\limsup_{t\to\infty} \frac{\left|\sum_{s=1}^t 1_{\{T_s=i\}} - m_i t\right|}{\sqrt{t\log\log t}} \leq C,$$

where for this ergodic Markov chain we use [3, Theorem 4, p. 90] to determine $m_i = \hat{\pi}_i$. The lemma follows easily. ∎

THEOREM 8.2. *Under* (H1)*, the sequence $\{I + \mathbf{a}_{\kappa_t} B_{T_t}, t \geq 0\}$ has ergodic backward products with probability one.*

Proof. Consider the consensus algorithm

$$Y_{t+1} = (I + \mathbf{a}_{\kappa_t} B_{T_t}) Y_t$$

with the deterministic initial pair (t_0, Y_{t_0}). Denote $\Lambda = \mathrm{diag}(\hat{\pi}_1^{-1}, \ldots, \hat{\pi}_n^{-1})$. We have

$$Y_{t+1} = (I + a_t \Lambda B_{T_t}) Y_t + (\mathbf{a}_{\kappa_t} - a_t \Lambda) B_{T_t} Y_t.$$

Select $t_1 \geq t_0$ such that $I + a_t \Lambda B_i$ is nonnegative for all $i \leq n$ and $t \geq t_1$. By Theorem 3.1, there exists a set N_1 with $P(N_1) = 0$ such that for all $\omega \in \Omega \backslash N_1$, $\{I + a_t \Lambda B_{T_t(\omega)}, t \geq t_1\}$ has ergodic backward products since $\cup_{i=1}^n \mathrm{graph}(\Lambda B_i) = G$ is strongly connected. Denote $Y_t = [Y_{t,1}, \ldots, Y_{t,n}]^T$. For some $C > 0$, we have a prior upper bound

$$|B_{T_t} Y_t| \leq C \max_j |Y_{t_0,j}|.$$

By Lemma 8.1, there exists a set N_2 with $P(N_2) = 0$ such that for all $\omega \in \Omega \backslash N_2$,

$$\begin{aligned}\left|a_{\kappa_t^i(\omega)} - a_t\hat{\pi}_i^{-\gamma}\right| &= O\left(\frac{1}{\left[t+O(\sqrt{t\log\log t})\right]^\gamma} - \frac{1}{t^\gamma}\right)\\ &= O\left(\frac{\sqrt{\log\log t}}{t^{\gamma+\frac{1}{2}}}\right).\end{aligned}$$

Note that both N_1 and N_2 are determined by $\{T_t, t \geq 0\}$. Denote $\Delta_t = (\mathbf{a}_{\kappa_t} - a_t\Lambda)B_{T_t}Y_t$. For each $\omega \in \Omega\backslash(N_1 \cup N_2)$, $\{I + a_t\Lambda B_{T_t(\omega)}, t \geq t_1\}$ has ergodic backward products and $\sum_{t=t_0}^\infty |\Delta_t(\omega)| < \infty$ since $\gamma \in (1/2, 1]$. By [9, Lemma B.2], there exists y such that $\lim_{t\to\infty} Y_t(\omega) = y1_n$ for $\omega \in \Omega\backslash(N_1 \cup N_2)$. Since (t_0, Y_{t_0}) can be arbitrarily selected, by [9, Lemma B.1], $\{I + \mathbf{a}_{\kappa_t(\omega)}B_{T_t(\omega)}, t \geq 0\}$ has ergodic backward products for all $\omega \in \Omega\backslash(N_1 \cup N_2)$. The theorem follows. ∎

THEOREM 8.3. *Under* (H1)–(H3)*, the algorithm* (8.2) *ensures mean square consensus.*

Proof. Since $\{T_t, t \geq 0\}$ is independent of $\{W_t, t \geq 0\}$, we have

$$E|\mathbf{a}_{\kappa_t}D_{T_t}W_t|^2 \leq CE|\mathbf{a}_{\kappa_t}|^2 = C\sum_{i=1}^n Ea_{\kappa_t^i}^2.$$

Fix i and as in the proof of Lemma 8.1, define the sequence $\{\tau_k, k \geq 1\}$. Set $\tau_0 = 0$.

Take a large $l_0 > 1$ so that $\{a_t, t \geq \tau_{l_0}\}$ satisfies $a_t = \frac{1}{t^\gamma}$. We have

$$\sum_{t=0}^\infty Ea_{\kappa_t^i}^2 = E\sum_{l=0}^\infty \sum_{t=\tau_l}^{\tau_{l+1}-1} a_{\kappa_t^i}^2.$$

Then for $l \geq l_0$, by (8.3)

$$E\sum_{t=\tau_l}^{\tau_{l+1}-1} a_{\kappa_t^i}^2 \leq \frac{E(\tau_{l+1}-\tau_l)}{(k_0^i+l)^{2\gamma}} \leq \frac{C}{(k_0^i+l)^{2\gamma}},$$

where $C > 0$ does not depend on l. By (8.3), it is easy to show

$$E\sum_{l=0}^{l_0-1}\sum_{t=\tau_l}^{\tau_{l+1}-1} a_{\kappa_t^i}^2 < \infty.$$

Consequently, $\sum_{t=0}^\infty Ea_{\kappa_t^i}^2 < \infty$ for each i, which implies that

$$\sum_{t=0}^\infty E|\mathbf{a}_{\kappa_t}D_{T_t}W_t|^2 < \infty. \tag{8.4}$$

By Theorem 8.2, (8.4), and (H1)–(H3), we apply [9, Theorem 3] to conclude that (8.2) ensures mean square consensus. ∎

9. Concluding remarks. We have studied ergodicity of backward products of a class of stochastic matrices with Markovian switches and decreasing step sizes. The ergodicity theorem is used to prove mean square consensus of stochastic approximation algorithms. Our proof of the ergodicity theorem assumes that the Markov chain is ergodic. An interesting question is what happens if the Markov chain is irreducible but periodic. This scenario seems to be more challenging. The dimension reduction technique via the canonical form in section 6 cannot be applied since the matrix P_π in this case has several eigenvalues with absolute value equal to one. To handle this scenario, a promising method is to explore the stochastic averaging approach [17] by identifying a limiting ordinary differential equation governing the stochastic approximation algorithm since the irreducible and periodic case still offers good long-run average properties for the model. We hope to pursue this idea in our future studies.

Appendix A.

Proof of Lemma 5.1. Associate $P_\pi = (q_{lm})_{l,m\le N}$ with a Markov chain $\{\theta'_t, t \ge 0\}$, whose irreducibility follows from that of $\{\theta_t, t \ge 0\}$. Since P_θ is ergodic, there exists $k_0 \ge 1$ such that for all $k \ge k_0$, the k-step transition probability $p_{11}^{[k]} > 0$. It implies that there exists a transition path $1, l_1, l_2, \dots, l_{k-1}, 1$ such that $p_{1l_1}p_{l_1l_2}\dots p_{l_{k-1}1} > 0$. For the Markov chain $\{\theta'_t, t \ge 0\}$, the probability of the path $1, l_{k-1}, \dots, l_2, l_1, 1$ is

$$q_{1l_{k-1}} \dots q_{l_2l_1}q_{l_11} = p_{l_{k-1}1}(\pi_{l_{k-1}}/\pi_1)\dots p_{l_1l_2}(\pi_{l_1}/\pi_{l_2})p_{1l_1}(\pi_1/\pi_{l_1}) > 0.$$

The k-step transition probability $q_{11}^{[k]} \ge q_{1l_{k-1}} \dots q_{l_2l_1}q_{l_11} > 0$ for all $k \ge k_0$ and so $\{\theta'_t, t \ge 0\}$ is aperiodic.

Since P_π is ergodic, it has a unique stationary distribution. For any $m \le N$,

$$\sum_{l=1}^N \pi_l q_{lm} = \sum_{l=1}^N \pi_l\pi_m\pi_l^{-1}p_{ml} = \pi_m.$$

This verifies that $\pi = (\pi_1, \dots, \pi_N)$ is its stationary distribution. The lemma follows. ∎

Lemma A.1. *Suppose that B is a $k \times k$ matrix having zero row sums and non-negative off-diagonal entries. Denote $Q = I_k \otimes B + B \otimes I_k$. If* graph($B$) *contains a spanning tree with root $k_0 \in \{1, \dots, k\}$, then* graph(Q) *contains a spanning tree with root $k_0 \in \{1, \dots, k^2\}$.*

Proof. Without loss of generality, we take $k_0 = 1$. We introduce a sufficiently small $\tau > 0$ to define a stochastic matrix $I + \tau Q$ corresponding to a discrete time Markov chain with state space $\{1, 2, \dots, k^2\}$. Denote $B = (b_{ij})_{i,j\le k}$. We have the blockwise representation

$$I + \tau Q = (\delta_{ij}(I + \tau B) + b_{ij}\tau I)_{i,j\le k}\,.$$

Partition the states of the Markov chain into the sets $S'_i = \{(i-1)k+1, \dots, ik\}$, $i = 1, \dots, k$. The ith diagonal block of $I + \tau Q$ is $I + \tau B + b_{ii}\tau I$. Since graph($B$) contains a spanning tree with root 1, each state of S'_i other than $(i-1)k+1$ can reach $(i-1)k+1$ by a sequence of transitions staying within S'_i.

Now it suffices to show that $(i-1)k+1$ can reach state $k_0 = 1$ with positive probability for $i > 1$. Consider $i = 2$, and all other cases are similar. Since graph(B) contains a spanning tree, there exists a product of the form

$$b_{2i_1}b_{i_1i_2}\dots b_{i_l1} > 0$$

and we can ensure that $2, i_1, i_2, \dots, i_l, 1$ are different integers from $\{1, 2, \dots, k\}$. Then we can show that there is a positive probability for the k^2 state Markov chain to make the sequence of transitions

$$(2-1)k+1 \to (i_1-1)k+1 \to (i_2-1)k+1 \to \dots \to (i_l-1)k+1 \to 1,$$

and the corresponding probability is obtained from $I + \tau Q$ as $\tau^{l+1}(b_{2i_1}b_{i_1i_2}\dots b_{i_l1})$. ∎

Proof of Theorem 5.2. Without loss of generality, suppose that node $i_0 = 1$ is the root of $G_{\cup,\mathrm{tr}}$. Due to the particular structure of M_1 and Lemma 5.1, for each S_j, any two states can reach one another by a transition path within S_j. Denote the stochastic matrix $M_\epsilon = (\bar{p}_{ij})_{i,j\le Nn^2}$ for $0 < \epsilon \le \frac{1}{4\beta}$. It suffices to show that from each state $i \in \{2, \dots, n^2\}$, there exists a transition path of $\{\Upsilon_t, t \ge 0\}$ to give

$$\bar{p}_{ii_1}\bar{p}_{i_1 i_2}\dots\bar{p}_{i_r 1} > 0. \tag{A.1}$$

Denote $Q_l = I_n \otimes B_l + B_l \otimes I_n$. Then $M_\epsilon = M_1 + \epsilon M_2 = (q_{lm}(I_{n^2} + \epsilon Q_m))_{l,m\le N}$.

Step 1. Let $Q = \sum_{l=1}^N Q_l$. Since $\cup_{k=1}^N G_k$ has a spanning tree $G_{\cup,\mathrm{tr}}$ with node 1 being the root, graph$(\sum_{l=1}^N B_l)$ contains a spanning tree with node $1 \in \{1, \dots, n\}$ being its root. So graph(Q) contains a spanning tree with node $1 \in \{1, \dots, n^2\}$ as its root by Lemma A.1. Therefore, $I + \frac{\epsilon}{N}Q$ is a stochastic matrix of positive diagonal entries where state 1 is reachable from any state in $\{2, \dots, n^2\}$ with positive probability. Thus, the first column of $(I + \frac{\epsilon}{N}Q)^{n^2-1}$ has only positive entries since each state can transit to state 1 with at most $n^2 - 1$ steps.

Step 2. Consider the product

$$D := q_{1j_1}(I + \epsilon Q_{j_1})q_{j_1 j_2}(I + \epsilon Q_{j_2})\dots q_{j_s 1}(I + \epsilon Q_1).$$

Since $(q_{lm})_{l,m\le N}$ is irreducible by Lemma 5.1, there exists an integer K_0 depending on $(q_{lm})_{l,m\le N}$ such that the above product has at most K_0 matrix terms, i.e., $s+1 \le K_0$, including each matrix in $\{Q_l, l \le N\}$ at least once and satisfying $q_{1j_1}q_{j_1 j_2}\dots q_{j_s 1} > 0$. For two nonnegative matrices, $A_1 \ge A_2$ means that the inequality holds componentwise. Note that $I + \epsilon Q_j \ge I/2$ since $0 < \epsilon \le \frac{1}{4\beta}$. Then $I + \epsilon Q_j \ge I/4 + I/2 + (\epsilon/2)Q_j$. For some constants C_1, C_2, we have the estimate

$$\begin{aligned} D &\ge C_1(I + \epsilon Q_1)\dots(I + \epsilon Q_N) \\ &\ge C_1\,[I/4 + I/2 + (\epsilon/2)Q_1]\dots[I/4 + I/2 + (\epsilon/2)Q_N] \\ &\ge C_1\left[4^{-N}I + 4^{-N+1}\sum_{l=1}^N (I/2 + (\epsilon/2)Q_l)\right] \\ &\ge C_2\left(I + \frac{\epsilon}{N}Q\right). \end{aligned}$$

So

$$D^{n^2-1} \ge C_2^{n^2-1}\left(I + \frac{\epsilon}{N}Q\right)^{n^2-1},$$

where the first column of $(I + \frac{\epsilon}{N}Q)^{n^2-1}$ has n^2 positive entries by Step 1. Thus, we may find a product of the form

$$D' := q_{1j_1}(I + \epsilon Q_{j_1})q_{j_1 j_2}(I + \epsilon Q_{j_2})\dots q_{j_{s'} 1}(I + \epsilon Q_1)$$

so that the first column has all positive entries.

Step 3. Take any $1 \le j \le n^2$. We check the $(j, 1)$th entry of D'. By Step 2,

$$\begin{aligned} D'(j,1) = \sum_{t_1,t_2,\dots,t_{s'}} & q_{1j_1}(I + \epsilon Q_{j_1})(j, t_1)q_{j_1 j_2}(I + \epsilon Q_{j_2})(t_1, t_2) \\ & \times \dots q_{j_{s'} 1}(I + \epsilon Q_1)(t_{s'}, 1) > 0. \end{aligned} \tag{A.2}$$

Recall that $M(i,j)$ denotes the (i,j)th entry of a matrix M. By (A.2), there exists a particular choice $(\hat{t}_1, \hat{t}_2, \dots, \hat{t}_{s'})$ such that

$$q_{1j_1}(I+\epsilon Q_{j_1})(j,\hat{t}_1)q_{j_1j_2}(I+\epsilon Q_{j_2})(\hat{t}_1,\hat{t}_2)\dots q_{j_{s'}1}(I+\epsilon Q_1)(\hat{t}_{s'},1)>0,$$

which implies that $\{\Upsilon_t, t\geq 0\}$ has the transition path

$$j \to (j_1-1)n^2+\hat{t}_1 \to (j_2-1)n^2+\hat{t}_2 \to \cdots \to (j_{s'}-1)n^2+\hat{t}_{s'} \to 1$$

with positive probability. Since $j\leq n^2$ is arbitrary, (A.1) holds. This completes the proof. ∎

LEMMA A.2. *Let $\{a_t, t\geq 1\}$ be a nonnegative sequence converging to zero (not necessarily satisfying* (A1)*). Suppose*

$$v_{t+1} = (I+a_tM)v_t + O(a_t^2), \quad t\geq 1,$$

where M is a Hurwitz matrix with all its eigenvalues having a real part strictly less than $-\sigma_0$ for some $\sigma_0>0$. Suppose that the sequence $\{b_t, t\geq 1\}$ satisfies $0<b_t\to 0$, $(I+a_tM)\frac{b_t}{b_{t+1}} = I+a_tM_0+o(a_t)$ for some Hurwitz matrix M_0, $\frac{a_t}{b_t}=O(1)$. Then the following assertions hold:

(i) $|v_t| = O(b_t)$.

(ii) *If $a_t = t^{-\gamma}$, $0<\gamma<1$, we have $|v_t|=O(t^{-\gamma})$. If $a_t=t^{-1}$, $|v_t|=O(t^{-\eta})$, where $\eta=\min\{1,\sigma_0\}$.*

Proof. We have

$$b_{t+1}^{-1}v_{t+1} = (I+a_tM_0+o(a_t))\left(b_t^{-1}v_t\right)+O(a_t), \quad t\geq 1.$$

Denote $r_t = b_t^{-1}v_t$. Taking any $Q>0$, we solve a unique $P>0$ from $PM_0+M_0^TP = -Q$. Then

$$\begin{aligned} r_{t+1}^TPr_{t+1} &= r_t^T(I+a_tM_0+o(a_t))^TP(I+a_tM_0+o(a_t))r_t+O(a_t^2)\\ &\quad + 2r_t^T(I+a_tM_0+o(a_t))^TPO(a_t), \end{aligned} \tag{A.3}$$

where $o(a_t)$ and $O(a_t)$ on the right-hand side of (A.3) are a matrix and a vector, respectively. Denote $d_t = r_t^TPr_t$. By taking a large t_0, we can find $\delta_0>0$ and $C_0>0$ to ensure

$$d_{t+1}\leq (1-\delta_0a_t)d_t + C_0a_t^2 + C_0a_t|r_t|, \quad t\geq t_0,$$

where $1-\delta_0a_t>0$. Next, we can find a large C_1 to ensure

$$C_0|r_t|\leq \frac{\delta_0}{2}d_t + C_1.$$

Hence for some $C_2>0$,

$$d_{t+1}\leq \left(1-\frac{\delta_0}{2}a_t\right)d_t + C_2a_t, \quad t\geq t_0.$$

Consider

$$h_{t+1} = \left(1-\frac{\delta_0}{2}a_t\right)h_t + C_2a_t, \quad h_{t_0}=d_{t_0}.$$

By induction we can show $0 \le d_t \le h_t$. On the other hand, it is easy to show that $h_t - \frac{2C_2}{\delta_0}$ converges to a finite limit $(h_{t_0} - \frac{2C_2}{\delta_0})\prod_{s=t_0}^{\infty}(1-\frac{\delta_0}{2}a_s)$. Hence $d_t = O(1)$. Part (i) follows.

Case 1: $a_t = t^{-\gamma}$, $0<\gamma<1$. We take $b_t = a_t$. It can be checked that

$$\frac{a_t}{a_{t+1}} = \left(1+t^{-1}\right)^{\gamma} = 1+\gamma t^{-1} + o\left(t^{-1}\right) = 1+o(a_t).$$

For this case $M_0 = M$.

Case 2: $a_t = t^{-1}$. We take $b_t = t^{-\eta}$. Then

$$(I+a_tM)\frac{b_t}{b_{t+1}} = (I+t^{-1}M)\left(1+\eta t^{-1}+O(t^{-2})\right) = I+t^{-1}(M+\eta I)+o(a_t).$$

The matrix $M_0 = M+\eta I$ is Hurwitz. Moreover, $\frac{a_t}{b_t} = O(1)$. This completes the proof of part (ii). ∎

Appendix B.

Proof of Theorem 3.1. Note that (5.4) is obtained from (4.6) by reordering the Nn^2 agents. By Theorem 6.4, S2 holds and hence S1 holds by Lemma 4.3.

Step 1. Consider any given deterministic value X_{t_0} for (4.1). There exists $\alpha \in \mathbb{R}$ such that

$$\lim_{t\to\infty} \bar{\xi}_t = \alpha 1_{Nn^2}.$$

Hence

$$\lim_{t\to\infty} \xi_t = \alpha \mathrm{diag}(\pi_1 I_{n^2}, \ldots, \pi_N I_{n^2}) 1_{Nn^2}. \tag{B.1}$$

For $V(t) = \sum_{l=1}^{N} V_l(t)$ and $J_n = \frac{1}{n}1_n 1_n^T$, (B.1) implies that $\lim_{t\to\infty} V(t) = \alpha 1_n 1_n^T = \alpha n J_n$. Next,

$$\begin{aligned} E|(I_n - J_n)X_t|^2 &= E\left[X_t^T (I_n - J_n)^2 X_t\right] \\ &= \mathrm{Tr}[(I_n - J_n)V(t)]. \end{aligned}$$

It is clear that

$$\lim_{t\to\infty} E|X_t - J_nX_t|^2 = 0,$$

which implies that the difference between the states of any two agents converges to zero in mean square. By the proving argument in [10, Theorem 9], we can further show mean square consensus of (4.1).

Step 2. Let $\{e_i, 1 \le i \le n\}$ be the canonical basis of $\mathbb{R}^n$. We set $X_{t_0} = X_{t_0}^{(i)} = e_i$, respectively, and by Step 1 we can show that

$$\begin{aligned} \Psi_{t+1,t_0} &:= (I+a_tB_{\theta_t})\ldots(I+a_{t_0}B_{\theta_{t_0}}) \\ &= (I+a_tB_{\theta_t})\ldots(I+a_{t_0}B_{\theta_{t_0}})\left[X_{t_0}^{(1)},\ldots,X_{t_0}^{(n)}\right] \\ &= \left[X_{t+1}^{(1)},\ldots,X_{t+1}^{(n)}\right] \end{aligned} \tag{B.2}$$

converges in mean square to a stochastic matrix of identical rows. By the method in [9, Theorem 3, necessity proof], we may further obtain that Ψ_{t,t_0} converges with probability one to a stochastic matrix of identical rows for the given t_0. This completes the proof of Theorem 3.1. ∎

Acknowledgments. The first author (M.H.) would like to express his gratitude to Professor Lei Guo and Professor Yiguang Hong of the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, for their warm hospitality during his visit.

REFERENCES

[1] T. C. Aysal and K. E. Barner, *Convergence of consensus models with stochastic disturbances*, IEEE Trans. Inform. Theory, 56 (2010), pp. 4101–4113.

[2] S. Chatterjee and E. Seneta, *Towards consensus: Some convergence theorems on repeated averaging*, J. Appl. Probab., 14 (1977), pp. 89–97.

[3] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin, 1960.

[4] J. E. Cohen, J. Hajnal, and C. M. Newman, *Approaching consensus can be delicate when positions harden*, Stochastic Process. Appl., 22 (1986), pp. 315–322.

[5] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-Time Markov Jump Linear Systems*, Springer-Verlag, London, 2005.

[6] F. Cucker and E. Mordecki, *Flocking in noisy environments*, J. Math. Pures Appl., 89 (2008), pp. 278–296.

[7] L. Elsner, I. Koltracht, and M. Neumann, *On the convergence of asynchronous paracontractions with applications to tomographic reconstruction from incomplete data*, Linear Algebra Appl., 130 (1990), pp. 65–82.

[8] F. Fagnani and S. Zampieri, *Randomize consensus algorithms over large-scale networks*, IEEE J. Selected Areas Commun., 26 (2008), pp. 639–649.

[9] M. Huang, *Stochastic approximation for consensus: A new approach via ergodic backward products*, IEEE Trans. Automat. Control, 57 (2012), pp. 2994–3008.

[10] M. Huang, S. Dey, G. N. Nair, and J. H. Manton, *Stochastic consensus over noisy networks with Markovian and arbitrary switches*, Automatica, 46 (2010), pp. 1571–1583.

[11] M. Huang and J. H. Manton, *Stochastic approximation for consensus seeking: Mean square and almost sure convergence*, in Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, 2007, pp. 306–311.

[12] M. Huang and J. H. Manton, *Stochastic consensus seeking with noisy and directed interagent communication: Fixed and randomly varying topologies*, IEEE Trans. Automat. Control, 55 (2010), pp. 235–241.

[13] S. Kar and J. M. F. Moura, *Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise*, IEEE Trans. Signal Process., 57 (2009), pp. 355–369.

[14] S. Kar, J. M. F. Moura, and H. V. Poor, *Distributed linear parameter estimation: Asymptotically efficient adaptive strategies*, SIAM J. Control Optim., 51 (2013), pp. 2200–2229.

[15] A. Kashyap, T. Basar, and R. Srikant, *Quantized consensus*, Automatica, 43 (2007), pp. 1192–1203.

[16] V. Krishnamurthy, K. Topley, and G. Yin, *Consensus formation in a two-time-scale Markovian system*, SIAM J. Multiscale Model. Simul., 7 (2009), pp. 1898–1927.

[17] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.

[18] T. Li and J.-F. Zhang, *Mean square average-consensus under measurement noises and fixed topologies*, Automatica, 45 (2009), pp. 1929–1936.

[19] T. Li and J.-F. Zhang, *Consensus conditions of multi-agent systems with time-varying topologies and stochastic communication noises*, IEEE Trans. Automat. Control, 55 (2010), pp. 2043–2057.

[20] I. Matei, J. S. Baras, and C. Somarakis, *Convergence results for the linear consensus problem under Markovian random graphs*, SIAM J. Control Optim., 51 (2013), pp. 1574–1591.

[21] I. Matei, N. Martins, and J. S. Baras, *Almost sure convergence to consensus in Markovian random graphs*, in Proceedings of the 47th IEEE Conference on Decision and Control, Cancun, Mexico, 2008, pp. 3535–3540.

[22] M. Nourian, P. E. Caines, R. P. Malhamé, and M. Huang, *Mean field LQG control in leader-follower stochastic multi-agent systems: Likelihood ratio based adaptation*, IEEE Trans. Automat. Control, 57 (2012), pp. 2801–2816.

[23] R. Rajagopal and M. J. Wainwright, *Network-based consensus averaging with general noisy channels*, IEEE Trans. Signal Process., 59 (2011), pp. 373–385.

[24] W. Ren, R. W. Beard, and D. B. Kingston, *Multi-agent Kalman consensus with relative uncertainty*, in Proceedings of American Control Conference, Portland, OR, 2005, pp. 1865–1870.

[25] E. Seneta, *Non-negative Matrices and Markov Chains*, 2nd ed., Springer, New York, 2006.

[26] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, *Decentralized parameter estimation by consensus based stochastic approximation*, IEEE Trans. Automat. Control, 56 (2011), pp. 531–543.

[27] B. Touri and A. Nedic, *On ergodicity, infinite flow and consensus in random models*, IEEE Trans. Automat. Control, 56 (2011), pp. 1593–1605.

[28] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.

[29] J. Wolfowitz, *Products of indecomposable, aperiodic, stochastic matrices*, Proc. Amer. Math. Soc., 14 (1963), pp. 733–737.

[30] L. Xiao, S. Boyd, and S.-J. Kim, *Distributed average consensus with least-mean-square deviation*, J. Parallel Distrib. Comput., 67 (2007), pp. 33–46.

[31] J. Xu, H. Zhang, and L. Xie, *Stochastic approximation approach for consensus and convergence rate analysis of multiagent systems*, IEEE Trans. Automat. Control, 57 (2012), pp. 3163–3168.

[32] Y. Yang and R. S. Blum, *Broadcast-based consensus with non-zero-mean stochastic perturbations*, IEEE Trans. Inform. Theory, 59 (2013), pp. 3971–3989.

[33] G. Yin, Y. Sun, and L. Y. Wang, *Asymptotic properties of consensus-type algorithms for networked systems with regime-switching topologies*, Automatica, 47 (2011), pp. 1366–1378.

[34] G. Yin, Q. Yuan, and L. Y. Wang, *Asynchronous stochastic approximation algorithms for networked systems: Regime-switching topologies and multi-scale structure*, SIAM J. Multiscale Model. Simul., 11 (2013), pp. 813–839.