Stochastic Approximation for Consensus: A New Approach via Ergodic Backward Products

Minyi Huang, *Member, IEEE*

***Abstract*—This paper considers both synchronous and asynchronous consensus algorithms with noisy measurements. For stochastic approximation based consensus algorithms, the existing convergence analysis with dynamic topologies heavily relies on quadratic Lyapunov functions, whose existence, however, may be difficult to ensure for switching directed graphs. Our main contribution is to introduce a new analytical approach. We first show a fundamental role of ergodic backward products for mean square consensus in algorithms with additive noise. Subsequently, we develop ergodicity results for backward products of degenerating stochastic matrices converging to a 0–1 matrix via a discrete time dynamical system approach. These results complement the existing ergodic theory of stochastic matrices and provide an effective tool for analyzing consensus problems. Under a joint connectivity assumption, our approach may deal with switching topologies, delayed measurements and correlated noises, and it does not require the double stochasticity condition typically assumed for the existence of quadratic Lyapunov functions.**

***Index Terms*—Backward product, consensus, delay, ergodicity, mean square convergence, measurement noise, stochastic approximation, synchronous and asynchronous algorithms.**

I. Introduction

During the past decade, an enormous amount of research effort has been devoted to consensus problems and various closely related formulations for multi-agent systems [8], [16], [20], [23], [32], [36]. A comprehensive survey may be found in [31], [37]. In recent years, consensus algorithms with imperfect information exchange have attracted significant attention, addressing additive noise [2], [3], [12], [30], [34], [38], [39], [46], [47] or quantization effect [4], [10], [11], [22]. The classic work [44] considered consensus algorithms for distributed function optimization with noisy gradient information. Models with noisy measurements take into account random uncertainty in signal reception and characterize more realistic network conditions. Also, when probabilistic quantizers are used to eliminate bias, quantization errors may be modeled as additive noise of zero mean [4].

For consensus models with noisy measurements, stochastic approximation with a decreasing step size may be applied such that each agent can gradually reduce the weights to its neighbors and hence attenuate the noise [18], [21], [35], [41]. A typical algorithm takes the form

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

where X_t is a vector consisting of the states of n nodes; $B_t = (b_{ij}(t))_{1\leq i,j\leq n}$ is a matrix with zero row sums and nonnegative off-diagonal entries; $a_t > 0$ is the step size; and W_t is a noise vector with coefficient matrix D_t. For $i \neq j$, $b_{ij}(t)$ is nonzero if and only if there is an edge from node j to node i at time t.

For fixed network topologies, B_t may be selected as a constant matrix B, and mean square and probability one convergence of the algorithm may be proved by either quadratic Lyapunov functions [18], [27] or change of coordinates [19].

Convergence analysis in switching networks has heavily relied on quadratic Lyaponuv functions [3], [17], [21], [28], [42], and an assumption often used is that the weight matrix $I + a_t B_t$ is doubly stochastic for each t (or starting from some t_0). The double stochasticity assumption was initially introduced for noiseless average-consensus problems to make the state average invariant [32]. For algorithm (1), this assumption ensures that the squared norm of the disagreement vector [28], [32] becomes a quadratic Lyapunov function. However, it is also very restrictive [5] and brings about serious feasibility issues in directed networks. To elucidate this aspect, let the $n \times n$ nonnegative matrix B_t^o be the same as B_t except for all zero diagonal entries. We call B_t^o a weight balanced matrix if the ith row sum equals the ith column sum for each i. Suppose that $I + a_t B_t$ is a nonnegative matrix and $a_t > 0$; then obviously $I + a_t B_t$ is doubly stochastic if and only if B_t^o is weight balanced, and it is known that such a matrix B_t^o exists if and only if the directed graph modeling the network is strongly semiconnected [15] (i.e., if there exists a directed path from node i to node j, then there exists one from j to i). In randomly varying directed networks, it is demanding to maintain strong semiconnectedness, and so one cannot in general expect the existence of doubly stochastic weight matrices at all times.

In addition to the above nonexistence issue, the double stochasticity requirement also imposes significant computational difficulties. For undirected graphs, doubly stochastic weight matrices may be constructed online via Metropolis weights [46]. However, on time-varying directed graphs it is generally infeasible to construct such matrices online without global instantaneous network topology information even if they exist. Although distributed iterative algorithms are available for constructing doubly stochastic matrices over directed graphs [15], they are not applicable for dynamic topologies. So, it is

Manuscript received May 07, 2010; revised January 19, 2011, January 20, 2011, and September 19, 2011; accepted April 03, 2012. Date of publication May 14, 2012; date of current version November 21, 2012. This work was supported in part by a Discovery Grant of the Natural Sciences and Engineering Research Council (NSERC) of Canada. A preliminary version of this paper was presented at the 49th IEEE CDC Conference, 2010. Recommended by Associate Editor C. Szepesvari.

The author is with the School of Mathematics and Statistics, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: mhuang@math.carleton.ca).

Digital Object Identifier 10.1109/TAC.2012.2199149

of practical importance to study models using general weight matrices without double stochasticity.

The recent work [3] considers the consensus algorithm

$$X_{t+1} = A_t X_t + F_t W_t, \quad t \geq 0 \tag{2}$$

where A_t is a random matrix of unit row sums and W_t is the additive noise. This model has considerable generality with respect to the past research; in particular, A_t is not required to be nonnegative. The authors succeeded in developing Lyapunov analysis without double stochasticity conditions. They proved probability one convergence of the algorithm and estimated the convergence rate. Let $J = (1/n)\mathbf{1}_{n\times n}$, where $\mathbf{1}_{n\times n}$ is an $n \times n$ matrix of all ones. The convergence proof in [3] uses $|X_t - JX_t|^2$ as a stochastic Lyapunov function and requires i) $\lambda_1\{E[A_t^T(I-J)A_t]\} \leq 1$, where $\lambda_1(\cdot)$ denotes the largest eigenvalue of a symmetric matrix, and ii) $\sum_{t=0}^{\infty}(1 - \lambda_1\{E[A_t^T(I-J)A_t]\}) = \infty$, in addition to assumptions on the noise term. Although this approach achieves considerable modeling generality, the eigenvalue conditions i) and ii) are quite restrictive for application to stochastic approximation as illustrated by the following example.

Example: Let $\{A_t, t \geq 0\}$ be a deterministic sequence. Each A_t takes a value from $\{I + a_t B^{(1)}, I + a_t B^{(2)}\}$, where $a_t = 1/(50 + t^{\gamma})$ for some $\gamma \in (1/2, 1]$ and

$$B^{(1)} = \begin{bmatrix} -24 & 0 & 24 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad B^{(2)} = \begin{bmatrix} -48 & 0 & 48 \\ 1 & -2 & 1 \\ 0 & 2 & -2 \end{bmatrix}.$$

The sequence $\{W_t, t \geq 0\}$ consists of independent random vectors, $EW_t = 0$, $\sup_{t\geq 0} E|W_t|^2 < \infty$, $E|X_0|^2 < \infty$, and $F_t = a_t I$. ♢

By the structure of $B^{(1)}$ and $B^{(2)}$, the network underlying algorithm (2) is always strongly connected. However, the criteria in [3] cannot be used to assert convergence. Denote the characteristic polynomial $f^{(k)}(\lambda) = \det[\lambda I - (I + a_t B^{(k)})^T (I - J)(I + a_t B^{(k)})]$, $k = 1, 2$. Then $\lim_{\lambda\to\infty} f^{(k)}(\lambda) = \infty$, and by direct calculations, $f^{(1)}(1) = -(88/3)a_t^2 + O(a_t^3)$, $f^{(2)}(1) = -(172/3)a_t^2 + O(a_t^3)$. For sufficiently small a_t, $f^{(k)}(\lambda) = 0$ has a root in $(1, \infty)$ by the intermediate value theorem; so $\lambda_1[A_t^T(I-J)A_t] > 1$ and conditions i)–ii) are violated.

To overcome the inherent limitations of the existing methods, we develop a new approach to analyze stochastic approximation for consensus, and it will cover many practical models and the above example as well. In a very general setup of noisy consensus algorithms, we show that ergodicity of the backward products of the weight matrices is a necessary and sufficient condition for mean square consensus, which reveals the fundamental mechanism governing the convergent algorithms studied by different authors [3], [17], [21], [28], [42]. By checking ergodicity of backward products in switching models, one may be saved from the challenging task of searching for Lyapunov functions. For nonexistence of quadratic Lyapunov functions with general weight matrices, see [33].

The weight matrices in our stochastic approximation algorithm, starting with an appropriate initial time to ensure nonnegative matrices, form a sequence of stochastic matrices converging to a 0–1 matrix, which will be called degenerating stochastic matrices due to this convergence feature. We note that there has existed an extensive literature (see [40] and references therein) on ergodicity of backward products of stochastic matrices. In particular, for analyzing inhomogeneous backward products, Wolfowitz's ergodicity theorem [20], [36], [45] and paracontraction [13], [25], [26] are well known powerful tools. Also, for the iterations of a nonnegative matrix with stationary delays, multiplicative ergodicity was studied via Lyaponov exponents in [14]. However, these results are not applicable to our model. To our best knowledge, this paper is the first to systematically establish ergodic theorems for backward products of degenerating stochastic matrices. We introduce a key notion of compatible nonnegative matrices and develop a dynamical system approach to prove ergodicity, which differs from the classical approaches in [40], [45].

Our approach to analyze the stochastic approximation algorithm differs from the Lyapunov approach by a shift of attention from ensuring steady energy decay to ergodicity check. This strategy makes it possible to overcome the weakness of the existing methods. Also, we note that this paper only addresses mean square convergence of the stochastic approximation algorithm. The further analysis of its sample path behavior is an interesting problem.

The main contributions of the paper are summarized as follows:

- We show a fundamental role of ergodic backward products for mean square consensus, and establish ergodic theorems for degenerating stochastic matrices.
- The ergodicity approach is applied to prove mean square consensus for a large class of models which the existing Lyapunov approaches cannot handle; specifically, it enables us to treat general weight matrices over switching networks, delayed noisy measurements, synchronous and asynchronous algorithms, and also correlated noises.

The organization of the paper is as follows. Section II formulates the stochastic consensus problem. Section III shows a necessary and sufficient condition for mean square consensus via ergodic backward products. Section IV introduces compatible nonnegative matrices and Section V proves ergodicity of degenerating stochastic matrices. Section VI shows mean square consensus, and Section VII concludes the paper.

We make some convention about notation. The node index is often used as a superscript in different variables (x_t^i, z_t^i, etc.) and should not be interpreted as an exponent of a number. For a matrix M, the element in the ith row and the jth column is called the (i,j)th element and denoted by $M(i,j)$. For a vector or matrix M, denote the Frobenius norm $|M| = [\mathrm{Tr}(M^T M)]^{1/2}$. Let $\mathbf{1}_k \in \mathbb{R}^k$ denote a column vector of k ones. For column vectors $Z_1, \ldots, Z_l$, $[Z_1; \ldots; Z_l]$ denotes the column vector obtained by vertical concatenation of the l vectors. For two sets S_1 and S_2, the set $S_1 \setminus S_2$ consists of all elements which are in S_1 but not in S_2. The abbreviation w.p.1 stands for "with probability one." We use C (or $C_0, C_1, \ldots$) to denote a generic positive constant which may vary at different places.

II. Stochastic Consensus Algorithm

For the reader's convenience, we provide a list of key notation used in the analysis.[1]

G	Digraph with nodes $\mathcal{N}=\{1,\dots,n\}$ and edges $\mathcal{E}$.
A_G	Adjacency matrix of G
G_t	Random digraph with nodes $\mathcal{N}$ and edges $\mathcal{E}_t$.
$G_{[t_1,t_2)}$	Union of the digraphs G_t, $t_1 \le t < t_2$.
x_t^i	State of node i.
y_t^{ik}	Noisy measurement node i obtains from node k.
w_t^{ik}	Additive noise in y_t^{ik}.
d_t^{ik}	Measurement delay along edge (k,i).
d^*	Deterministic upper bound of all d_t^{ik}.
d_1^*	d^*+1.
θ_t	Counters $(\theta_t^1,\dots,\theta_t^n)$ of n nodes.
a_t	Step size for state update.
$b_{ik}(t)$	Weight parameter node i assigns to node k.
B_t	$n\times n$ matrix $(b_{ik}(t))_{1\le i,k\le n}$ of zero row sums.
$\mathbf{B}_t$	$nd_1^*\times nd_1^*$ matrix.
$\mathbf{U}$	$nd_1^*\times nd_1^*$ stochastic matrix of 0–1 elements.
$\mathbf{D}_t$	$nd_1^*\times n_1$ noise coefficient matrix.
$X_t, X_t^{(-d)}$	n-dimensional vector.
$\mathbf{X}_t$	nd_1^*-dimensional vector.
W_t	n_1-dimensional noise vector.
Z_t	n-dimensional vector obtained by re-ordering entries in X_t.
$\mathbf{A}_t$	$nd_1^*\times nd_1^*$ stochastic matrix.
M_t, M_t^A	Stochastic matrices.
$\Phi_{t,s}, M^{t,s}$	Backward product of stochastic matrices.

We introduce some standard preliminaries on graph modeling of the network topology. A directed graph (or 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 $i_1,\dots,i_l$, $l\ge 2$, such that $(i_k,i_{k+1})\in\mathcal{E}$ for all $k\le l-1$. 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. The adjacency matrix of G is an $n\times n$ matrix $A_G=(a_{ij})_{1\le i,j\le n}$, where $a_{ij}=1$ if $(i,j)\in\mathcal{E}$, and $a_{ij}=0$ otherwise.

The dynamic network topology to specify the signal reception is modeled by a sequence of digraphs $\{G_t=(\mathcal{N},\mathcal{E}_t),t\ge 0\}$, where $\mathcal{N}=\{1,\dots,n\}$ and $\mathcal{E}_t$ randomly changes with time. We may view $\{\mathcal{E}_t,t\ge 0\}$ as a set-valued random process. The adjacency matrix A_{G_t} is a matrix-valued random variable. So $\mathcal{E}_t$ is completely determined by A_{G_t}. If $(j,i)\in\mathcal{E}_t$, node i receives information from node j which is called a neighbor of node i. The neighbor set of node i is denoted by $\mathcal{N}_{i,t}=\{j|(j,i)\in\mathcal{E}_t\}$.

[1]Letters in boldface usually denote vectors or matrices built upon more basic ones in lower dimensions.

A. Stochastic Approximation Algorithm

Let the underlying probability space be denoted by $(\Omega,\mathcal{F},P)$, corresponding to the sample space, the collection of all events, and the probability measure, respectively. At time $t\in\{0,1,2,\dots\}$, node i is associated with a real-valued state x_t^i. Each node knows its own state exactly. Define the state vector

$$X_t=\left[x_t^1,\dots,x_t^n\right]^T,\quad t\ge 0.$$

The initial state vector is X_0. At time t, if $\mathcal{N}_{i,t}\neq\emptyset$ (the empty set), node i receives possibly outdated information from its neighbors, which is modeled by

$$y_t^{ik}=x_{t-d_t^{ik}}^k+w_t^{ik},\quad k\in\mathcal{N}_{i,t} \tag{3}$$

where w_t^{ik} is the noise and $d_t^{ik}\ge 0$ is an integer-valued random delay. Since the system starts at $t=0$, the implicit requirement for the neighbor set is that

$$k\in\mathcal{N}_{i,t}\quad\text{implies}\quad t-d_t^{ik}\ge 0. \tag{4}$$

A fixed upper bound for d_t^{ik} will be specified later. Each node will use its own state and its noisy measurements to form a weighted average.

According to the local information exchange, we define the matrix $B_t=(b_{ik}(t))_{1\le i,k\le n}$ as follows. If $\mathcal{N}_{i,t}=\emptyset$, define

$$b_{ik}(t)=0\quad\text{for all } k\in\mathcal{N}. \tag{5}$$

If $\mathcal{N}_{i,t}\neq\emptyset$, set

$$\begin{cases} b_{ik}(t)\in[\underline{b},\overline{b}], & \text{if } k\in\mathcal{N}_{i,t}\\ b_{ik}(t)=0, & \text{if } k\notin\mathcal{N}_{i,t}\cup\{i\}\\ b_{ii}(t)=-\sum_{k\in\mathcal{N}_{i,t}}b_{ik}(t) \end{cases} \tag{6}$$

where $0<\underline{b}\le\overline{b}<\infty$ are two deterministic constants. We may interpret $B(t)$ as the generator of a continuous time Markov chain with n states. The specification (5), (6) is a generalization of the weights with a fixed network [19] to randomly varying networks, and $\{B_t,t\ge 0\}$ is a matrix-valued random process.

Each node maintains a counter θ_t^i. Denote $\theta_t=[\theta_t^1,\dots,\theta_t^n]$. We describe two cases.

(SU) Synchronous update:

$$\theta_t^i=t,\quad i\in\mathcal{N},\quad t\ge 0 \tag{7}$$

where the nodes need to share slotted time.

(AU) Asynchronous update:

$$\theta_t^i=\sum_{s=1}^{t}1_{\{|\mathcal{N}_{i,s}|>0\}},\quad i\in\mathcal{N},\quad t\ge 1 \tag{8}$$

and $\theta_0^i=0$, where $|\mathcal{N}_{i,s}|$ is the number of neighbors of node i at time s. So (8) means that the node increases its counter by one whenever it receives signals from its neighbors.

The consensus algorithms for cases (SU) and (AU) are specified in a unified manner. At time $t \geq 0$, if $\mathcal{N}_{i,t} = \emptyset$, set $x_{t+1}^i \equiv x_t^i$. If $\mathcal{N}_{i,t} \neq \emptyset$, node i updates its state by the rule

$$x_{t+1}^i = \left[1 + a_{\theta_t^i} b_{ii}(t)\right] x_t^i + a_{\theta_t^i} \sum_{k \in \mathcal{N}_{i,t}} b_{ik}(t) y_t^{ik}, \quad t \geq 0 \quad (9)$$

where $\{a_t, t \geq 0\}$ is a sequence of positive step sizes. We call $I + \mathrm{Diag}(a_{\theta_t^1}, \ldots, a_{\theta_t^n}) B_t$ the weight matrix.

When the step size is updated using (8), the resulting asynchronous algorithm (9) is essentially driven by event times, i.e., the moments of receiving signals. Once initialized, the algorithm may be implemented without synchronized time slots although we use the pre-specified discrete times 0, 1, 2,... to describe (9). For related literature on asynchronous stochastic approximation, see [1], [7], [24], and [43].

The actually observed network topology at time t may be denoted by $G_t(\omega) = (\mathcal{N}, \mathcal{E}_t(\omega))$ to indicate its dependence on the sample. Denote the maximal set of communication links $\mathcal{E}_{\max} = \{(k,i) | P((k,i) \in \mathcal{E}_t) > 0 \text{ for some } t \geq 0\}$. For convenience of statistical modeling of the noises and delay, we make the convention: w_t^{ik} and d_t^{ik} are defined for all $(k,i) \in \mathcal{E}_{\max}$. If (k,i) does not appear in $\mathcal{E}_t$ so that the measurement relation (3) does not physically occur, we still introduce w_t^{ik} and d_t^{ik} as dummy random variables. Let the random variables $\{w_t^{ik} | (k,i) \in \mathcal{E}_{\max}\}$ be listed by a fixed ordering of (k,i) to obtain a noise vector W_t of n_1 dimension.

Definition 1: 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 $1 \leq i \leq n$. ♢

There have existed some effective approaches to analyze asynchronous stochastic approximation and show probability one convergence (see, e.g., [1], [43]). These algorithms are typically associated with an underlying time-invariant mapping, and the contraction property of the mapping may be exploited [43] or the ordinary differential equation (ODE) approach may be used after appropriate scaling of time [1] where the asynchronous algorithm behaves like a synchronous one with small perturbations. To analyze our model, due to the rapid switches of B_t, it is difficult to apply these approaches and it is necessary to develop a different method.

B. Main Assumptions

(**A1**) The deterministic sequence $\{a_t, t \geq 0\}$ satisfies

$$a_0 > 0, \quad \text{and} \quad \alpha t^{-\gamma} \leq a_t \leq \beta t^{-\gamma} \text{ for } t \geq 1 \qquad (10)$$

where $\gamma \in (1/2, 1]$ and $0 < \alpha \leq \beta < \infty$. ♢

So (**A1**) implies the standard step size condition used in stochastic approximation: $\sum_{t=0}^{\infty} a_t = \infty$ and $\sum_{t=0}^{\infty} a_t^2 < \infty$.

For two integers $0 \leq t_1 < t_2$, define the digraph $G_{[t_1,t_2)} = (\mathcal{N}, \cup_{t_1 \leq t < t_2} \mathcal{E}_t)$, which is called the union of the collection of digraphs $\{G_t | t_1 \leq t < t_2\}$. Since the sequence $\{G_t, t \geq 0\}$ depends on the sample ω, $G_{[t_1,t_2)}$ also depends on ω. We introduce the following assumption.

(**A2**) There exists an infinite sequence of integer-valued random variables $0 \equiv T_0 < T_1 < T_2 < \ldots$ such that the two conditions hold:

1) $G_{[T_l, T_{l+1})}$ is strongly connected w.p.1 for $l \geq 0$;
2) $\alpha_1 := \sup_{l \geq 1}(T_l - T_{l-1}) < \infty$ w.p.1. ♢

It will be helpful to provide some explanation on notation. Note that given a sample $\omega \in \Omega$, $G_t(\omega)$ has the set of edges $\mathcal{E}_t(\omega)$. The random digraph $G_{[T_l,T_{l+1})}$ has the set of nodes $\mathcal{N}$. Given ω, $G_{[T_l,T_{l+1})}$ is interpreted as $G_{[T_l(\omega),T_{l+1}(\omega))}(\omega)$ possessing the set of edges $\cup_{T_l(\omega) \leq t < T_{l+1}(\omega)} \mathcal{E}_t(\omega)$.

(**A3**) $\{W_t, t \geq 0\}$ is a sequence of independent random vectors of zero mean and is independent of $\{(B_t, A_{G_t}, \{d_t^{ik} | (k,i) \in \mathcal{E}_{\max}\}), t \geq 0\}$, where $0 \leq d_t^{ik} \leq d^*$ w.p.1 for a fixed integer d^*. In addition, $E|X_0|^2 < \infty$ and $\sup_{t \geq 0} E|W_t|^2 < \infty$. ♢

The delay upper bound d^* is used for analyzing the stochastic approximation algorithm, which may be implemented without knowing the value of d^*. To deal with leader following, we introduce another type of connectivity condition.

(**A2′**) There is a fixed leader node i_L which has no neighbor in each G_t. There exists an infinite sequence of integer-valued random variables $0 \equiv T_0 < T_1 < T_2 < \ldots$ such that w.p.1, $G_{[T_l,T_{l+1})}$ contains a spanning tree with root i_L for $l \geq 0$. In addition, (**A2**)-2) is satisfied. ♢

Under (**A2′**), there is no edge from other nodes to node i_L, so that B_t defined by (5), (6) necessarily has all zero elements on the i_Lth row.

C. Vector Form of the Algorithm

For $t \geq 0$, denote the set of $n \times n$ random matrices

$$B_{d,t} = (B_{d,t}(i,k))_{1 \leq i,k \leq n}, \quad d = 0, 1, \ldots, d^*.$$

For their diagonal elements, we take $B_{0,t}(i,i) = b_{ii}(t)$ and $B_{d,t}(i,i) = 0$, $d = 1, \ldots, d^*$, for all i. For $d = 0, 1, \ldots, d^*$, the off-diagonal element $B_{d,t}(i,k)$ is nonzero and further taken as $b_{ik}(t)$ if and only if $b_{ik}(t) > 0$ and $d_t^{ik} = d$. Denote

$$\mathbf{a}_{\theta_t}^0 = \mathrm{Diag}\left(a_{\theta_t^1}, \ldots, a_{\theta_t^n}\right), \quad \mathbf{B}_t^0 = [B_{0,t}, B_{1,t}, \ldots, B_{d^*,t}]. \qquad (11)$$

The ith row of $\mathbf{B}_t^0$ contains the same set of nonzero elements as the ith row of B_t does. Due to (4), if $t < d^*$, we necessarily have $B_{d,t} = 0$ for all $d = t+1, \ldots, d^*$.

We write (9) in the equivalent form

$$X_{t+1} = X_t + \mathbf{a}_{\theta_t}^0 \mathbf{B}_t^0 [X_t; X_{t-1}; \ldots; X_{t-d^*}] + \mathbf{a}_{\theta_t}^0 D_t W_t, \quad t \geq 0 \qquad (12)$$

where D_t is an $n \times n_1$ random matrix determined by B_t and we set $X_t \equiv 0$ for $-d^* \leq t < 0$. If $d^* = 0$ and all nodes update their step sizes by (7), then (12) reduces to (1).

For $d_1^* = d^* + 1$, denote the $nd_1^* \times nd_1^*$ matrix

$$\mathbf{U} = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix} \qquad (13)$$

where each identity matrix is $n \times n$, and denote $\mathbf{a}_{\theta_t} = \mathrm{Diag}(\mathbf{a}_{\theta_t}^0, 0_{nd^* \times nd^*})$,

$$\mathbf{B}_t = \begin{bmatrix} \mathbf{B_t^0} \\ 0_{nd^* \times nd_1^*} \end{bmatrix}, \quad \mathbf{D}_t = \begin{bmatrix} D_t \\ 0_{nd^* \times n_1} \end{bmatrix}. \qquad (14)$$

It is clear that $\mathbf{B}_t$ is determined by $(B_t, \{d_t^{ik} | (k,i) \in \mathcal{E}_{\max}\})$.

Denote $\mathbf{X}_t = [X_t; X_{t-1}; \ldots; X_{t-d^*}] \in \mathbb{R}^{nd_1^*}$. We have the state space representation

$$\mathbf{X}_{t+1} = (\mathbf{U} + \mathbf{a}_{\theta_t}\mathbf{B}_t)\,\mathbf{X}_t + \mathbf{a}_{\theta_t}\mathbf{D}_t W_t, \quad t \geq 0 \tag{15}$$

where $X_t \equiv 0$ for $-d^* \leq t < 0$. If all counters increase to infinity, there exists $t_0(\omega)$, depending on the sample ω, such that $\{\mathbf{U} + \mathbf{a}_{\theta_t(\omega)}\mathbf{B}_t(\omega), t \geq t_0(\omega)\}$ is a sequence of stochastic matrices converging to the 0–1 stochastic matrix $\mathbf{U}$. In this case, $\{\mathbf{U} + \mathbf{a}_{\theta_t(\omega)}\mathbf{B}_t(\omega), t \geq t_0(\omega)\}$ is called a sequence of degenerating stochastic matrices.

III. Necessary and Sufficient Condition for Consensus

We use a general algorithm to reveal a fundamental relationship between mean square consensus and ergodicity of backward matrix products. Consider the system

$$Y_{t+1} = A_t Y_t + H_t V_t, \quad t \geq 0 \tag{16}$$

where $Y_t \in \mathbb{R}^{m_1}$ denotes the states of m_1 agents, $V_t \in \mathbb{R}^{m_2}$ is the noise vector, and the initial condition is Y_0. Here $\{A_t, t \geq 0\}$ and $\{H_t, t \geq 0\}$ are two sequences of random matrices of compatible dimensions. For each fixed ω, $A_t(\omega)$ is a stochastic matrix for all $t \geq 0$. The model (16) includes (15) as a special case when the coefficient matrices of $\mathbf{X}_t$ in (15) are nonnegative for all $t \geq 0$.

A. Ergodicity of Backward Products

Let $\{\tilde{A}_t, t \geq 0\}$ be a sequence of deterministic nonnegative matrices, where each $\tilde{A}_t$ is a stochastic matrix. Define the so-called backward product

$$\Phi_{t,s} = \tilde{A}_{t-1} \ldots \tilde{A}_s \quad \text{for } t > s \geq 0, \quad \Phi_{s,s} := I.$$

The product $\Phi_{t,s}$ is still a stochastic matrix. Let $\Phi_{t,s}(i,j)$ denote its (i,j)th element.

Definition 2: [40] We say weak ergodicity holds for backward products of the sequence of stochastic matrices $\{\tilde{A}_t, t \geq 0\}$ if

$$\lim_{t\to\infty} |\Phi_{t,s}(i_1, j) - \Phi_{t,s}(i_2, j)| = 0$$

for any given $s \geq 0$ and i_1, i_2, j, i.e., the difference between any two rows of $\Phi_{t,s}$ converges to zero as $t \to \infty$. If in addition to weak ergodicity, $\Phi_{t,s}(i,j)$ converges as $t \to \infty$, for any s, i, j, we say strong ergodicity holds. ♢

By [40, p. 154, Th. 4.17], weak and strong ergodicity are equivalent for backward products of any sequences of stochastic matrices. Hence, in the following we only speak of ergodicity of backward products.

It is worth mentioning that weak and strong ergodicity may also be defined for forward products of the form $\Phi_{s,t}^f = \tilde{A}_s \ldots \tilde{A}_{t-1}$ for $t > s$, and $\Phi_{s,s}^f := I$, and that weak ergodicity differs from strong ergodicity. For an example showing divergence of the forward products of weakly ergodic stochastic matrices, see [9, p. 240].

B. Necessary and Sufficient Condition for Consensus

For the theorem below, we run algorithm (16) with any initial time-state pair (t_0, Y_{t_0}). Denote $Y_t = [Y_t^1, \ldots, Y_t^{m_1}]^T$, $\Psi_{t,s} = A_{t-1} \ldots A_s$ for $t > s$, and $\Psi_{s,s} := I$.

Theorem 3: Assume

i) $\{V_t, t \geq 0\}$ is a sequence of random vectors of zero mean, independent of $\{(A_t, H_t), t \geq 0\}$;
ii) $\sum_{t=0}^{\infty} E|H_t|^2 E|V_t|^2 < \infty$;
iii) there exists a sequence of nonnegative numbers $\{\phi_k, k \geq 0\}$ such that

$$\left|E\left[V_k V_{k'}^T\right]\right| \leq \phi_{|k-k'|}\left(E|V_k|^2 E|V_{k'}|^2\right)^{\frac{1}{2}}, \quad \sum_{k=0}^{\infty} \phi_k < \infty.$$

Then (16) ensures mean square consensus for any initial time-state pair (t_0, Y_{t_0}) with $E|Y_{t_0}|^2 < \infty$ if and only if $\{A_t, t \geq 0\}$ has ergodic backward products w.p.1. □

Remark: If the random vectors $\{V_t, t \geq 0\}$ are independent with $EV_t = 0$ and $E|V_t|^2 < \infty$, iii) holds with $\phi_0 = 1$ and $\phi_k = 0$ for all $k > 0$. ♢

The proof of Theorem 3 is given in Appendix A. In the context of stochastic approximation for consensus seeking, condition ii) in Theorem 3 is easy to satisfy since a decreasing step size may be used to attenuate the measurement noise. Thus, to a very large extent, the analysis of the asymptotic behavior of consensus algorithms of the form (16) reduces to checking the ergodicity condition along sample paths.

IV. Compatible Nonnegative Matrices

This section develops some basic tools for analyzing sequences of degenerating stochastic matrices. We first introduce a class of $nd_1^* \times nd_1^*$ stochastic matrix sequences motivated by the stochastic approximation based consensus algorithm with delay, and then show that their ergodicity analysis is equivalent to that of a sequence of $n \times n$ degenerating stochastic matrices. Our main idea of studying the backward products is to run a noiseless switching linear dynamical system governed by these matrices. By setting different initial conditions, the sequence of state vectors, as the output of the linear system, will reflect information on the backward products. The analysis of the state vectors is simpler than directly handling the backward products. To avoid introducing too many variables, the vectors X_t and $\mathbf{X_t}$ appearing in Section II will be reused in different systems and this should cause no risk of confusion.

A. Compatible Matrices

Let $\{\mathbf{A}_t, t \geq 0\}$ be $nd_1^* \times nd_1^*$ deterministic nonnegative matrices, where $d_1^* = d^* + 1$. Each $\mathbf{A}_t$ is a stochastic matrix of the form

$$\mathbf{A}_t = \begin{bmatrix} A_{00,t} & \ldots & A_{0d^*,t} \\ \vdots & \ddots & \vdots \\ A_{d^*0,t} & \ldots & A_{d^*d^*,t} \end{bmatrix} = \left[\begin{array}{ccc|c} A_{00,t}, & \ldots, & A_{0(d^*-1),t} & A_{0d^*,t} \\ \hline & I & & 0_{nd^*\times n} \end{array}\right] \tag{17}$$

where each matrix $A_{ik,t}$ is $n \times n$ and the identity matrix is $nd^* \times nd^*$. Let $\mathbf{A}_t(i,j)$ be the (i,j)th element of $\mathbf{A}_t$. We consider a class of stochastic matrix sequences motivated by the consensus algorithm (15). When t is large, each $\mathbf{A_t}$ is nearly a 0–1 matrix. However, these nearly zero elements still have an important effect on the asymptotic property of the backward products. Our idea is to introduce the notion of compatible matrices by associating $\mathbf{A}_t$ with a digraph $\tilde{G}_t$ where the latter identifies some relatively strong transitions described by $\mathbf{A}_t$. To facilitate further analysis, this compatibility notion is also defined for $n \times n$ matrices.

Definition 4: Let $\{\delta_t, t \geq 0\}$ be a sequence of nonnegative numbers with $\lim_{t\to\infty} \delta_t = 0$ and $\{\tilde{G}_t = (\mathcal{N}, \tilde{\mathcal{E}}_t), t \geq 0\}$ be a sequence of digraphs of n nodes.

a) The sequence of $nd_1^* \times nd_1^*$ stochastic matrices $\{\mathbf{A}_t, t \geq 0\}$ of the form (17) is said to be (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$ if there exist constants t_c, $0 < \underline{c} \leq \bar{c}$ such that for all $t \geq t_c$

$$\mathbf{A}_t(i,j) \leq \bar{c}\delta_t, \qquad \forall 1 \leq i \leq n, \quad 1 \leq j \leq nd_1^*, \quad i \neq j, \tag{18}$$

$$\max_{0\leq d\leq d^*} \mathbf{A}_t(i, j+dn) \geq \underline{c}\delta_t, \qquad \forall (j,i) \in \tilde{\mathcal{E}}_t. \tag{19}$$

b) The sequence of $n \times n$ stochastic matrices $\{M_t, t \geq 0\}$ is said to be (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$ if there exist t_c, $0 < \underline{c} \leq \bar{c}$ such that

$$M_t(i,j) \leq \bar{c}\delta_t, \qquad \forall 1 \leq i,j \leq n, \quad i \neq j, \tag{20}$$

$$M_t(i,j) \geq \underline{c}\delta_t, \qquad \forall (j,i) \in \tilde{\mathcal{E}}_t \tag{21}$$

where $M_t(i,j)$ is the (i,j)th element of M_t. ♢

Remark: We may further define (δ_t)-compatibility with any positive initial time t_0 in an obvious manner. If $d^* = 0$, part a) of Definition 4 reduces to part b). If $\{\mathbf{A}_t, t \geq 0\}$ is (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$, this property still holds if δ_t is replaced by $c\delta_t$ for some $c > 0$. ♢

Example: Take $\mathcal{N} = \{1,2\}$, $\mathcal{E}_0 = \{(2,1)\}$ and $\mathcal{E}_1 = \{(1,2),(2,1)\}$. Let $\{\tilde{G}_t, t \geq 1\}$ be defined by $\tilde{G}_{2k} = (\mathcal{N}, \mathcal{E}_0)$ and $\tilde{G}_{2k-1} = (\mathcal{N}, \mathcal{E}_1)$ for $k \geq 1$. Let

$$\mathbf{A}_{2k} = \begin{bmatrix} 1-\frac{1}{k} & 0 & \frac{3}{4k} & \frac{1}{4k} \\ \frac{1}{2k^2} & 1-\frac{1}{k^2} & \frac{1}{2k^2} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\mathbf{A}_{2k-1} = \begin{bmatrix} 1-\frac{1}{k} & \frac{1}{2k} & 0 & \frac{1}{2k} \\ \frac{1}{2k} & 1-\frac{1}{k} & \frac{1}{2k} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

It is easy to verify that $\{\mathbf{A}_t, t \geq 1\}$ is $(1/t)$-compatible with $\{\tilde{G}_t, t \geq 1\}$, where we take $\bar{c} = 2$ and $\underline{c} = 1/3$ for conditions (18), (19). For instance, when $t = 2k$, node 2 is a neighbor of node 1. In relation to (19), we check the first row of $\mathbf{A}_{2k}$, and the maximum of its second and fourth elements is $1/(4k) \geq \underline{c}/(2k)$. ♢

Example: Let $\{G_t(\omega), t \geq 0\}$ be given in Section II and $\{B_t(\omega), t \geq 0\}$ be specified by (5), (6), where $\omega \in \Omega$ is fixed. Then $\{\mathbf{A}_t(\omega) := \mathbf{U} + a_t\mathbf{B}_t(\omega), t \geq t_0(\omega)\}$ is (a_t)-compatible with $\{G_t(\omega), t \geq t_0(\omega)\}$ if $\{a_t, t \geq 0\}$ satisfies **(A1)** and $\mathbf{A}_t(\omega)$ is a nonnegative matrix for $t \geq t_0(\omega)$. ♢

Given $\mathbf{A}_t$ in (17), define

$$M_t^A = \sum_{d=0}^{d^*} A_{0d,t} \tag{22}$$

which is a stochastic matrix. The following property is obvious and the proof is omitted.

Proposition 5: Let $\{M_t^A, t \geq 0\}$ be defined by (22). If $\{\mathbf{A}_t, t \geq 0\}$ is (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$, then so is $\{M_t^A, t \geq 0\}$. □

The backward products of $\{\mathbf{A}_t, t \geq 0\}$ are closely related to the weighted averaging algorithm

$$\mathbf{X}_{t+1} = \mathbf{A}_t\mathbf{X}_t, \qquad t \geq 0. \tag{23}$$

To distinguish from the notation in (15), for (23) we denote

$$\mathbf{X}_t = \left[X_t; X_t^{(-1)}; \ldots; X_t^{(-d^*)}\right] \tag{24}$$

where X_t and $X_t^{(-d)}$, $1 \leq d \leq d^*$, are n dimensional. Further denote the n-dimensional linear system

$$X_{t+1} = M_t^A X_t, \qquad t \geq 0. \tag{25}$$

With the aid of the linear systems (23) and (25), we may prove the following equivalence theorem.

Theorem 6: Assume that $\{\mathbf{A}_t, t \geq 0\}$ is (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$ and $\sum_{t=0}^{\infty} \delta_t^2 < \infty$. Let $\{M_t^A, t \geq 0\}$ be defined by (22). Then $\{\mathbf{A}_t, t \geq 0\}$ has ergodic backward products if and only if $\{M_t^A, t \geq 0\}$ has ergodic backward products.

Proof: See Appendix B. □

The important implication of Theorem 6 is that the ergodicity property of the delay based sequence $\{\mathbf{A}_t, t \geq 0\}$ may be studied via a lower dimensional sequence without involving delay.

B. Linear Systems Governed by $n \times n$ Compatible Nonnegative Matrices

With Theorem 6 in mind, below we focus on analyzing $n \times n$ stochastic matrices. For a sequence of $n \times n$ stochastic matrices $\{M_t, t \geq 0\}$, consider the linear system

$$X_{t+1} = M_t X_t, \quad t \geq 0 \tag{26}$$

which may be interpreted as a consensus algorithm over a digraph $G_{M,t}$ of n nodes. The edges of $G_{M,t}$ are uniquely determined by the nonzero off-diagonal elements in M_t. We introduce the assumption.

(H1) i) $\{M_t, t \geq 0\}$ is (a_t)-compatible with a sequence of digraphs $\{\tilde{G}_t, t \geq 0\}$, where $\{a_t, t \geq 0\}$ satisfies **(A1)**; ii) each $\tilde{G}_t$ is strongly connected. ♢

Remark: $\tilde{G}_t$ may be different from $G_{M,t}$. ♢

C. State Reordering and Mutual Attraction of Trajectories

We run (26) with any fixed initial pair (t_0, X_{t_0}), where $t_0 \geq 0$. If X_t is considered directly for convergence analysis,

the problem is quite difficult since the components of X_t undergo very complex evolution. We define a new vector Z_t by reordering the n entries in $X_t = [x_t^1, \ldots, x_t^n]^T$. Suppose

$$x_t^{i_1} \geq x_t^{i_2} \geq \ldots \geq x_t^{i_n}$$

where $\{i_1, \ldots, i_n\}$ is a permutation of $\{1, \ldots, n\}$ and in general changes with time. We interpret x_t^i as the state of node i, $1 \leq i \leq n$. Define

$$Z_t = \left[z_t^1, \ldots, z_t^n\right]^T := \left[x_t^{i_1}, \ldots, x_t^{i_n}\right]^T. \tag{27}$$

After reordering of the states, the evolution of Z_t has more orderly behavior than that of X_t.

In later analysis, we need to identify a component of Z_t as a component of X_t and further check its recursive equation. We use the following rule to avoid indeterminacy when the same value repeats within Z_t. If there are exactly $l \geq 2$ components $z_t^k, \ldots, z_t^{k+l-1}$ within Z_t taking the same value c_l corresponding to l nodes with indices $i_1 < i_2 < \ldots < i_l$, the kth component z_t^k in Z_t is interpreted as the state of the node with the smallest index i_1; similarly, the $(k+1)$th component in Z_t is associated with the second smallest node index i_2, and so on. Define the n scalar sequences

$$\left\{z_t^k, t \geq t_0\right\}, \quad 1 \leq k \leq n. \tag{28}$$

We call $\{z_t^k, t \geq t_0\}$ the level-k trajectory. In analyzing $\{Z_t, t \geq t_0\}$, the basic idea is to show that there is a mutual attraction of the trajectories of different levels, which eventually merge to the same limit. Due to the degenerating stochastic matrices, our method differs from directly comparing the gap between the greatest and least states in a consensus algorithm with time varying weight matrices [6].

Since each M_t is a stochastic matrix, $\{z_t^1, t \geq t_0\}$ (resp., $\{z_t^n, t \geq t_0\}$) is bounded and non-increasing (resp., non-decreasing) (see, e.g., [6]), and so has a limit. We summarize this fact in the following proposition.

Proposition 7: Let Z_t be defined by (26), (27). Then both $\{z_t^1, t \geq t_0\}$ and $\{z_t^n, t \geq t_0\}$ converge to finite limits. □

The asymptotic behavior of the other sequences $\{z_t^k, t \geq t_0\}$, where $k \in \{2, \ldots, n-1\}$, is less obvious. The following theorem proves one of the key results of this paper. It is instrumental for establishing ergodicity of backward products of degenerating stochastic matrices. Its proof is quite involved. The basic idea is to use induction. First, the level-1 trajectory converges to a finite limit. Next, we show that each level-$(k+1)$ trajectory converges to the same limit as the level-k trajectory.

Theorem 8: Let (X_t, Z_t) be defined by (26), (27) with any initial condition X_{t_0}, $t_0 \geq 0$, and assume (**H1**). Then there exists $c \in \mathbb{R}$ such that $\lim_{t\to\infty} Z_t = \lim_{t\to\infty} X_t = c\mathbf{1}_n$.

Proof: See Appendix C. □

Remark: The proof of Theorem 8 only requires $\gamma \in (0, 1]$ for (10). The further restriction $\gamma \in (1/2, 1]$ is needed for proving mean square convergence in Section VI. ♢

D. Leader Following

A leader following structure may be incorporated into (26), and in this case (**H1**) is replaced by the following assumption.

(**H1′**) i) $\{M_t, t \geq 0\}$ is (a_t)-compatible with a sequence of digraphs $\{\tilde{G}_t, t \geq 0\}$, where $\{a_t, t \geq 0\}$ satisfies (**A1**); ii) $M(i_L, i_L) = 1$; iii) each $\tilde{G}_t$ contains a spanning tree, and all these spanning trees share a common root i_L which has no neighbor in $\tilde{G}_t$. ♢

(**H1′**)-ii) ensures that the i_Lth element of X_t is fixed as the leader' state $x_0^{i_L}$.

Our strategy of proving convergence for the leader following model is to start with a special class of initial conditions of X_t so that we may adapt the argument in proving Theorem 8 which has been based on jointly strongly connected digraphs. To treat general initial conditions, we apply a transformation of the initial condition so that the analysis is reduced to the previous case.

Corollary 9: Let X_t be defined by (26) with any initial condition X_{t_0}, $t_0 \geq 0$. Assuming (**H1′**), then $\lim_{t\to\infty} X_t = x_{t_0}^{i_L}\mathbf{1}_n$.

Proof: By (**H1′**)-ii), it follows that $x_t^{i_L} = x_{t_0}^{i_L}$ for $t \geq t_0$. Take a sufficiently large $t_1 \geq t_0 + 1$ such that by (20) after δ_t is replaced by a_t

$$\min_{1\leq i\leq n} M_t(i,i) > \frac{1}{2}, \qquad \forall t \geq t_1. \tag{29}$$

Now let X_{t_1} be generated with the initial pair (t_0, X_{t_0}) and used as the initial condition for

$$X_{t+1} = M_t X_t, \qquad t \geq t_1. \tag{30}$$

Proposition 7 still holds in the leader following case.

Step 1) Assume that for (30)

$$x_{t_0}^{i_L} = x_{t_1}^{i_L} < x_{t_1}^j, \quad \forall j \in S_F = \{1, \ldots, n\} \setminus \{i_L\}. \tag{31}$$

By (29) and induction it is straightforward to show that $x_t^j > x_{t_0}^{i_L}$ for $j \in S_F$ and all $t \geq t_1$. Hence, $x_t^{i_L} < x_t^j$ for $j \in S_F$ and $t \geq t_1$. Following (27), we still reorder the n components of X_t in (30) from the greatest to the least to obtain the vector Z_t, $t \geq t_1$. So $z_t^n = x_t^{i_L}$ for $t \geq t_1$.

First, by Proposition 7, $\{z_t^1, t \geq t_1\}$ has a finite limit c. Next, by using the same induction argument as in proving Theorem 8, we may show that all the n sequences $\{z_t^k, t \geq t_1\}$, $k = 1, \ldots, n$, must converge to the same limit c. In particular, when we apply induction, the argument for deriving (C.15) from (C.14) is still valid under (**H1′**)-iii) since there exists at least one edge pointing to the set of nodes associated with the first $l \leq n-1$ levels of trajectories from the remaining nodes which contain the leader since $z_t^n = x_t^{i_L}$. On the other hand, since $z_t^n \equiv x_{t_0}^{i_L}$, the limit must be $c = x_{t_0}^{i_L}$. We conclude that $\lim_{t\to\infty} Z_t = x_{t_0}^{i_L}\mathbf{1}_n$, and subsequently $\lim_{t\to\infty} X_t = x_{t_0}^{i_L}\mathbf{1}_n$.

Step 2) Consider X_{t_1} not satisfying the inequality in (31). Let $e_i \in \mathbb{R}^n$ be a unit vector with the ith element equal to 1. Select $C_L > 0$ such that the element at the i_Lth position of $Y_{t_1} = X_{t_1} - C_L e_{i_L}$ is less than any other element. Take $X'_{t_1} = Y_{t_1}$ as the initial condition of

$$X'_{t+1} = M_t X'_t, \qquad t \geq t_1. \tag{32}$$

By Step 1, $\lim_{t\to\infty} X'_t = (x^{i_L}_{t_0} - C_L)\mathbf{1}_n$. Take the initial condition $X''_{t_1} = -C_L e_{i_L}$ for

$$X''_{t+1} = M_t X''_t, \qquad t \geq t_1. \tag{33}$$

By Step 1, $\lim_{t\to\infty} X''_t = -C_L\mathbf{1}_n$. By linearity of (32), (33), $X_t = X'_t - X''_t$, $t \geq t_1$, where X_t is generated by (30). Hence, $\lim_{t\to\infty} X_t = x^{i_L}_{t_0}\mathbf{1}_n$. □

V. Ergodicity of Degenerating Stochastic Matrices

Our plan of proving ergodicity is as follows. For the simple case where $\{M_t, t \geq 0\}$ in (26) is compatible with a sequence of strongly connected digraphs $\{\tilde{G}_t, t \geq 0\}$, the convergence result in Theorem 8 is applicable. For the general case with joint connectivity, our strategy is to reduce to the simple case. We select a sequence of times $0 =: \tau_0 < \tau_1 < \dots$ to form a union of digraphs and a backward sub-product on each subinterval $[\tau_i, \tau_{i+1})$, where each union of digraphs is strongly connected. This is a typical procedure to exploit joint connectivity; see, e.g., [20] and [36]. Then the original backward product may be written as the product of these sub-products, possibly together with extra multiplicative terms at two ends. A key step is to show that the sub-products have desired properties to generate compatible stochastic matrices (see Lemmas D.1 and D.2).

Throughout Sections V-A and V-B, all matrices and digraphs are deterministic.

A. Ergodicity of Backward Products

For the sequence $\{M_t, t \geq 0\}$, define the backward product

$$\Phi_{t,s} = M_{t-1}\dots M_s, \quad t > s \geq 0$$

and $\Phi_{s,s} := I$. The following ergodicity theorem is an easy consequence of Theorem 8.

Theorem 10: Assuming (**H1**), ergodicity holds for the backward products of $\{M_t, t \geq 0\}$.

Proof: Let $e_i \in \mathbb{R}^n$ be the unit column vector with the ith element equal to 1. For any fixed $s \geq 0$, by Theorem 8 there exists $\pi_i(s)$ depending on s such that $\lim_{t\to\infty}\Phi_{t,s}e_i = \pi_i(s)\mathbf{1}_n$. By Lemma B.1, the theorem follows. □

To generalize Theorem 10, we consider ergodicity for stochastic matrices associated with jointly strongly connected digraphs. For a sequence of digraphs $\{\tilde{G}_t = (\mathcal{N}, \tilde{\mathcal{E}}_t), t \geq 0\}$, we follow the rule in Section II to define the union of digraphs. For two integers $0 \leq \tau_t < \tau_{t+1}$, define

$$\widehat{M}_t = M_{\tau_{t+1}-1}\dots M_{\tau_t}, \quad \widehat{G}_t = \tilde{G}_{[\tau_t,\tau_{t+1})}. \tag{34}$$

Theorem 11: Assume i) $\{M_t, t \geq 0\}$ is (a_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$, where $\{a_t, t \geq 0\}$ satisfies (**A1**); ii) the sequence $0 =: \tau_0 < \tau_1 < \dots$ satisfies $\sup_{i\geq 0}(\tau_{i+1} - \tau_i) < \infty$; and iii) $\tilde{G}_{[\tau_i,\tau_{i+1})}$ is strongly connected for each $i \geq 0$. Then ergodicity holds for the backward products of $\{M_t, t \geq 0\}$.

Proof: By Lemma D.2 and Theorem 10, the backward products of $\{\widehat{M}_t, t \geq 0\}$ are ergodic. For any $s \geq 0$, there exists some $i_0 \geq 1$ such that $\tau_{i_0-1} \leq s < \tau_{i_0}$. Denote $\Pi_0 = \lim_{t\to\infty}\widehat{M}_t\dots\widehat{M}_{i_0}$, which is a stochastic matrix of identical rows. For any $\epsilon > 0$, there exists $K_0 > i_0$ such that

$$|\widehat{M}_l\dots\widehat{M}_{i_0} - \Pi_0| \leq \epsilon$$

for all $l \geq K_0$. Now for all $t \geq \tau_{K_0+1} + 1$, we have

$$\begin{aligned}
&M_{t-1}\dots M_s \\
&\quad = M_{t-1}\dots M_{\tau_{K_0+1}}\widehat{M}_{K_0}\dots\widehat{M}_{i_0}M_{\tau_{i_0}-1}\dots M_s \\
&\quad = M_{t-1}\dots M_{\tau_{K_0+1}}\left[\widehat{M}_{K_0}\dots\widehat{M}_{i_0} - \Pi_0\right]M_{\tau_{i_0}-1}\dots M_s \\
&\qquad + M_{t-1}\dots M_{\tau_{K_0+1}}\Pi_0 M_{\tau_{i_0}-1}\dots M_s \\
&\quad = M_{t-1}\dots M_{\tau_{K_0+1}}\left[\widehat{M}_{K_0}\dots\widehat{M}_{i_0} - \Pi_0\right]M_{\tau_{i_0}-1}\dots M_s \\
&\qquad + \Pi_0 M_{\tau_{i_0}-1}\dots M_s
\end{aligned}$$

where $\Pi_0 M_{\tau_{i_0}-1}\dots M_s$ is a stochastic matrix of identical rows. Since $M_{t-1}\dots M_{\tau_{K_0+1}}$ and $M_{\tau_{i_0}-1}\dots M_s$ are stochastic matrices, $|M_{t-1}\dots M_{\tau_{K_0+1}}[\widehat{M}_{K_0}\dots\widehat{M}_{i_0} - \Pi_0]M_{\tau_{i_0}-1}\dots M_s| \leq n\epsilon$. Since $\epsilon > 0$ is arbitrary, $\lim_{t\to\infty}\Phi_{t,s} = \Pi_0 M_{\tau_{i_0}-1}\dots M_s$.

Remark: Theorems 10 and 11 hold with $\gamma \in (0, 1]$ for (10) since the proofs of Theorem 8 and Lemma D.1 only need $\gamma \in (0, 1]$. ♢

B. Backward Products of $\{\mathbf{A}_t, t \geq 0\}$

Theorem 12: Assume i) $\{\mathbf{A}_t, t \geq 0\}$ is (a_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$, where $\{a_t, t \geq 0\}$ satisfies (**A1**); ii) the sequence $0 =: \tau_0 < \tau_1 < \dots$ satisfies $\sup_{i\geq 0}(\tau_{i+1} - \tau_i) < \infty$; and iii) $\tilde{G}_{[\tau_i,\tau_{i+1})}$ is strongly connected for each $i \geq 0$. Then ergodicity holds for the backward products of $\{\mathbf{A}_t, t \geq 0\}$.

Proof: By Proposition 5, $\{M^A_t, t \geq 0\}$ is (a_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$, and so has ergodic backward products by Theorem 11. By Theorem 6, $\{\mathbf{A}_t, t \geq 0\}$ has ergodic backward products. □

C. Application to Random Networks

In (**A2**), we may select a null set N_0 (i.e., $P(N_0) = 0$) such that $G_{[T_l(\omega),T_{l+1}(\omega))}(\omega)$ is strongly connected and $\alpha_1(\omega) = \sup_{l\geq 0}|T_{l+1}(\omega) - T_l(\omega)| < \infty$ for $\omega \in \Omega \setminus N_0$.

Corollary 13: Assume i) $\{B_t, t \geq 1\}$ is given by (5), (6) and (**A1**)–(**A2**) hold; ii) $s_0(\omega)$ is an integer such that each $\mathbf{A}^\dagger_t = \mathbf{U} + \mathbf{a}_{\theta_t}\mathbf{B}_t$, $t \geq s_0(\omega)$, is a stochastic matrix. Then for each $\omega \in \Omega \setminus N_0$, ergodicity holds for the backward products of $\{\mathbf{A}^\dagger_t(\omega), t \geq s_0(\omega)\}$.

Proof: For synchronous step size update, $\{\mathbf{A}^\dagger_t(\omega), t \geq s_0(\omega)\}$ is (a_t)-compatible with $\{G_t(\omega), t \geq s_0(\omega)\}$.

We further check the compatibility condition with asynchronous step size update. Take $\omega \in \Omega \setminus N_0$. We have $1 \leq \alpha_1(\omega) < \infty$. Consider any node i_0 and integer $l \geq 1$. Since $G_{[T_l(\omega),T_{l+1}(\omega))}(\omega)$ is strongly connected, there exists at least one node j_0 as a neighbor of node i_0 in $G_{[T_l(\omega),T_{l+1}(\omega))}(\omega)$. Suppose that (j_0, i_0) is an edge in $G_{T_l(\omega)+\xi}(\omega)$, where $0 \leq \xi < T_{l+1}(\omega) - T_l(\omega)$. Then

$$\theta^{i_0}_{T_l(\omega)-1}(\omega) + 1 \leq \theta^{i_0}_{T_{l+1}(\omega)-1}(\omega) \leq \theta^{i_0}_{T_l(\omega)-1}(\omega) + \alpha_1(\omega) \tag{35}$$

where the second inequality follows from $\theta^{i_0}_{k+j} \leq \theta^{i_0}_k + j$ for $j \geq 1$. Since $\{\theta^{i_0}_t, t \geq 0\}$ is non-decreasing, by (35) there exist coefficients $0 < c_1(\omega) \leq c_2(\omega)$ such that

$$c_1(\omega)t \leq \theta^{i_0}_t(\omega) \leq c_2(\omega)t$$

for all $t \geq T_1(\omega)$. Subsequently,

$$\alpha\left[c_2(\omega)t\right]^{-\gamma} \leq a_{\theta_t^{i_0}(\omega)} \leq \beta\left[c_1(\omega)t\right]^{-\gamma}, \quad t \geq T_1(\omega). \quad (36)$$

Next, for $\omega \in \Omega \setminus N_0$, by (36) we may check that $\{\mathbf{A}_t^\dagger(\omega), t \geq s_0(\omega)\}$ is (a_t)-compatible with $\{G_t(\omega), t \geq s_0(\omega)\}$. The corollary follows from Theorem 12. □

VI. Mean Square Consensus

Let (15) be rewritten in the form

$$\mathbf{X}_{t+1} = \mathbf{A}_t^\dagger \mathbf{X}_t + \mathbf{a}_{\theta_t}\mathbf{D}_t W_t, \qquad t \geq 0$$

where $\mathbf{A}_t^\dagger = \mathbf{U} + \mathbf{a}_{\theta_t}\mathbf{B}_t$. By (9), (12) and (14), $\mathbf{D}_t$ is a function of B_t, and $\sup_{t\geq 0,\omega\in\Omega}|\mathbf{D}_t| < \infty$.

Theorem 14: Assuming (**A1**)–(**A3**), mean square consensus holds for (9) with synchronous step size update, i.e., $\lim_{t\to\infty} E|x_t^i - x^*|^2 = 0$ for some x^* and for all $1 \leq i \leq n$.

Proof: We have $\mathbf{a}_{\theta_t} = \mathrm{Diag}(a_t I_n, 0_{nd^*\times nd^*})$ and $\{W_t, t \geq 0\}$ is independent of $\{(\mathbf{A}_t^\dagger, \mathbf{D}_t), t \geq 0\}$. Take a sufficiently large t_0 such that w.p.1 $\mathbf{A}_t^\dagger$ is a stochastic matrix for $t \geq t_0$. In addition,

$$\sum_{t=0}^{\infty} E\left|\mathbf{a}_{\theta_t}\mathbf{D}_t\right|^2 E|W_t|^2 = \sum_{t=0}^{\infty} a_t^2 E|D_t|^2 E|W_t|^2 < \infty.$$

The theorem follows from Theorem 3 and Corollary 13. □

Theorem 15: Assume (**A1**)–(**A3**). In addition, i) there exists a deterministic integer t_0 such that w.p.1 $\mathbf{A}_t^\dagger$ is a nonnegative matrix for all $t \geq t_0$; ii) $E\sup_{l\geq 1}\Delta_l^{2\gamma} < \infty$, where $\Delta_l = T_{l+1} - T_l$. Then algorithm (9) with asynchronous step size update ensures mean square consensus.

Proof: For the asynchronous case, $\mathbf{a}_{\theta_t}$ is a matrix function of the adjacency matrices $(A_{G_k}, 0 \leq k \leq t)$. It is easy to see that $\{W_t, t \geq 0\}$ is independent of $\{(\mathbf{A}_t^\dagger, \mathbf{a}_{\theta_t}\mathbf{D}_t), t \geq 0\}$. For some constant C, we have

$$\sum_{t=0}^{\infty} E|\mathbf{a}_{\theta_t}\mathbf{D}_t|^2 E|W_t|^2 \leq C\sum_{t=0}^{\infty} E\left|\mathbf{a}_{\theta_t}\right|^2 = C\sum_{t=0}^{\infty}\sum_{i=1}^{n} E\left|a_{\theta_t^i}\right|^2.$$

For some constant C_0, $a_{\theta_t^i} \leq C_0 \sup_{l\geq 1}\Delta_l^\gamma / t^\gamma$ for $t \geq 1$ since $\theta_t^i \geq (t/\sup_{l\geq 1}\Delta_l) - 1$. Subsequently,

$$\sum_{t=0}^{\infty} E\left|\mathbf{a}_{\theta_t}\mathbf{D}_t\right|^2 E|W_t|^2 \leq C_1 a_0^2 + C_1 E\sup_{l\geq 1}\Delta_l^{2\gamma}\sum_{t=1}^{\infty} t^{-2\gamma} < \infty.$$

The theorem follows from Theorem 3 and Corollary 13. □

Remark: If $\sup_{l\geq 1}\Delta_l \leq C$ w.p.1 for some constant C, conditions i) and ii) in Theorem 15 hold. ♢

For leader following, if node i_L is the leader, the i_Lth row of D_t has all zeros for each t due to (9). The convergence for the leader following case may be proved by an ergodicity approach based on Corollary 13. The proof is omitted.

Corollary 16: In Theorems 14 and 15, if (**A2**) is replaced by (**A2′**) while other assumptions still hold, then $\lim_{t\to\infty} E|x_t^i - x_0^{i_L}|^2 = 0$ for all i. □

Remark: Theorems 14–15 and Corollary 16 may be generalized to correlated noises by letting $\{W_t, t \geq 0\}$ satisfy condition iii) in Theorem 3. ♢

Remark: If (3) is replaced by $y_t^{ik} = (x_s^k + w_s^{ik})|_{s=t-d_t^{ik}}$, Theorems 14–15 and Corollary 16 still hold. ♢

VII. Concluding Remarks

We considered synchronous and asynchronous stochastic approximation for consensus seeking with delayed measurements in dynamic noisy environments. This paper developed ergodicity results for degenerating stochastic matrices and proved mean square consensus without quadratic Lyapunov functions. In future work, it will be of interest to relax the bounded time interval condition for joint connectivity so that the modeling may deal with more general random networks, such as Markovian switching networks [29], [17]. Convergence rate bounds and probability one convergence of the consensus algorithm are also interesting topics.

Appendix A
Proof of Theorem 3

Lemma A.1: Denote $\xi_{T,j} = \sum_{k=T}^{T+j-1}\Psi_{T+j,k+1}H_k V_k$ for $T \geq 0$ and $j \geq 1$. Then

$$E|\xi_{T,j}|^2 \leq 2m_1 m_2 \sum_{l=0}^{\infty}\phi_l \sum_{k=T}^{T+j-1} E|H_k|^2 E|V_k|^2.$$

Proof: For all ω and $t \geq s$, $\Psi := \Psi_{t,s}(\omega)$ is a stochastic matrix ensuring $|\Psi|^2 = \mathrm{Tr}(\Psi^T\Psi) = \mathrm{Tr}(\Psi\Psi^T) \leq m_1$. We have

$$E|\xi_{T,j}|^2 = \sum_{T\leq i,k\leq T+j-1} E\left[V_i^T H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k V_k\right]$$

For each pair (i,k),

$$\begin{aligned}
&\left|E\left[V_i^T H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k V_k\right]\right| \\
&= \left|\mathrm{Tr}\left\{E\left[H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k\right] E\left[V_k V_i^T\right]\right\}\right| \\
&\leq m_2\left|E\left[H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k\right] E\left[V_k V_i^T\right]\right| \\
&\leq m_2 E\left|H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k\right| \cdot \left|E\left[V_k V_i^T\right]\right| \\
&\leq m_1 m_2 \phi_{|k-i|} E\left(\left|H_i^T\right| \cdot |H_k|\right)\left(E|V_k|^2 E|V_i|^2\right)^{\frac{1}{2}} \\
&\leq m_1 m_2 \phi_{|k-i|}\left(E|H_i|^2 E|H_k|^2\right)^{\frac{1}{2}}\left(E|V_k|^2 E|V_i|^2\right)^{\frac{1}{2}} \\
&\leq \left(\frac{m_1 m_2}{2}\right)\phi_{|k-i|}\left(E|H_i|^2 E|V_i|^2 + E|H_k|^2 E|V_k|^2\right). \quad (A.1)
\end{aligned}$$

Hence, by (A.1)

$$\left|\sum_{T\leq i,k\leq T+j-1} E\left[V_i^T H_i^T \Psi_{T+j,i+1}^T \Psi_{T+j,k+1} H_k V_k\right]\right|$$

$$\le \left(\frac{m_1 m_2}{2}\right) \sum_{l=-j+1}^{j-1} \phi_{|l|} \sum_{k=T}^{T+j-1} 2E|H_k|^2 E|V_k|^2$$
$$\le 2m_1 m_2 \sum_{l=0}^{\infty} \phi_l \sum_{k=T}^{T+j-1} E|H_k|^2 E|V_k|^2.$$

The lemma follows. □

Proof of Theorem 3: Sufficiency—For fixed (t_0, Y_{t_0})

$$Y_{t+1} = \Psi_{t+1,t_0} Y_{t_0} + \sum_{k=t_0}^{t} \Psi_{t+1,k+1} H_k V_k, \quad t \ge t_0.$$

By Lemma A.1, we may show that $\sup_{t\ge t_0} E|Y_t|^2 < \infty$.

Let $\xi_{T,j} = \sum_{k=T}^{T+j-1} \Psi_{T+j,k+1} H_k V_k$ for $T \ge t_0$ and $j \ge 1$. For any given $\epsilon > 0$, by Lemma A.1 we may select $t_1 \ge t_0$ such that

$$\sup_{T\ge t_1} \sup_{j\ge 1} E|\xi_{T,j}|^2 \le \frac{\epsilon}{2}. \tag{A.2}$$

Taking a fixed $T' \ge t_1$, we have

$$Y_{t+1} = \Psi_{t+1,T'} Y_{T'} + \sum_{k=T'}^{t} \Psi_{t+1,k+1} H_k V_k, \quad t \ge T'.$$

Since $\Psi_{t+1,T'}$ converges w.p.1 to a stochastic matrix $\Pi_{T'}$ of identical rows, there exists a random variable $\zeta_{T'}$ such that

$$\lim_{t\to\infty} E\left|\Psi_{t+1,T'} Y_{T'} - \zeta_{T'} \mathbf{1}_{m_1}\right|^2 = 0. \tag{A.3}$$

So by (A.2), (A.3)

$$\limsup_{t\to\infty} E\left|Y_{t+1}^i - Y_{t+1}^j\right|^2 \le 2 \limsup_{k\to\infty} E|\xi_{T',k}|^2 \le \epsilon$$

and therefore, there exists $t_2 \ge t_1$ such that for all i, j and $t \ge t_2$

$$E\left|Y_t^i - Y_t^j\right|^2 \le 2\epsilon. \tag{A.4}$$

Following the argument in [7, Th. 9], we proceed to show mean square consensus. For any $t' > t \ge t_2$, we have $Y_{t'} = \Psi_{t',t} Y_t + \xi_{t,t'-t}$ and

$$\begin{aligned}
&E|Y_{t'} - Y_t|^2 \\
&\le 2E|\Psi_{t',t} Y_t - Y_t|^2 + 2E|\xi_{t,t'-t}|^2 \\
&\le 2E\left|\left(\Psi_{t',t} Y_t^1 \mathbf{1}_{m_1} - Y_t\right) + \Psi_{t',t}\left(Y_t - Y_t^1 \mathbf{1}_{m_1}\right)\right|^2 + \epsilon \\
&\le 4E\left|Y_t^1 \mathbf{1}_{m_1} - Y_t\right|^2 + 4E\left(|\Psi_{t',t}|^2 \left|Y_t - Y_t^1 \mathbf{1}_{m_1}\right|^2\right) + \epsilon \\
&\le 4m_1(2\epsilon) + 4m_1(2m_1\epsilon) + \epsilon = \left(8m_1^2 + 8m_1 + 1\right)\epsilon.
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, $\{Y_t, t \ge t_0\}$ is a Cauchy sequence in the L_2 norm and has a limit $Y_\infty = [Y_\infty^1, \ldots, Y_\infty^{m_1}]^T$. Furthermore, $E|Y_\infty^i - Y_\infty^j|^2 = \lim_{t\to\infty} E|Y_t^i - Y_t^j|^2 = 0$ by arbitrariness of $\epsilon > 0$ in (A.4). So $Y_\infty = Y_\infty^1 \mathbf{1}_{m_1}$ and mean square consensus follows.

Necessity—Given any $\epsilon > 0$ and $t_0 \ge 0$, we select $t_1 \ge t_0$ such that (A.2) holds. Let $e_i \in \mathbb{R}^{m_1}$ be a unit column vector with the ith element equal to 1. Take the initial pair (t_1, Y_{t_1}) for (16) with $Y_{t_1} = e_i$. We have

$$Y_{t+1} = \Psi_{t+1,t_1} e_i + \sum_{k=t_1}^{t} \Psi_{t+1,k+1} H_k V_k, \quad t \ge t_1.$$

By mean square consensus there exists a random variable η_i such that

$$\lim_{t\to\infty} E\left|Y_{t+1} - \eta_i \mathbf{1}_{m_1}\right|^2 = 0. \tag{A.5}$$

We have

$$\begin{aligned}
E\left|\Psi_{t+1,t_1} e_i - \eta_i \mathbf{1}_{m_1}\right|^2 &\le 2E\left|Y_{t+1} - \eta_i \mathbf{1}_{m_1}\right|^2 \\
&\quad + 2E\left|\xi_{t_1,t-t_1+1}\right|^2 \\
&\le 2E\left|Y_{t+1} - \eta_i \mathbf{1}_{m_1}\right|^2 + \epsilon.
\end{aligned}$$

This combined with (A.5) implies that there exists $t_2 \ge t_1$ such that for $t \ge t_2$ and each e_i,$1 \le i \le m_1$, $E|\Psi_{t+1,t_1} e_i - \eta_i \mathbf{1}_{m_1}|^2 \le 2\epsilon$. Since $\Psi_{t+1,t_1} e_i$ is the ith column of Ψ_{t+1,t_1}, it follows that

$$E\left|\Psi_{t+1,t_1} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\right|^2 \le 2m_1\epsilon, \quad t \ge t_2.$$

Here we will not directly analyze the asymptotic property of Ψ_{t+1,t_1} since t_1 changes with ϵ. In fact $(\eta_1, \ldots, \eta_{m_1})$ also changes with t_1 and ϵ. For the fixed t_0, we have

$$\begin{aligned}
&\left|\Psi_{t+1,t_0} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\Psi_{t_1,t_0}\right|^2 \\
&= \left|\Psi_{t+1,t_1}\Psi_{t_1,t_0} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\Psi_{t_1,t_0}\right|^2 \\
&\le \left|\Psi_{t+1,t_1} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\right|^2 \left|\Psi_{t_1,t_0}\right|^2 \\
&\le m_1 \left|\Psi_{t+1,t_1} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\right|^2.
\end{aligned}$$

Hence,

$$E\left|\Psi_{t+1,t_0} - \mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\Psi_{t_1,t_0}\right|^2 \le 2m_1^2\epsilon \tag{A.6}$$

where $t \ge t_2$. Since $\epsilon > 0$ is arbitrary, (A.6) implies that $\{\Psi_{t+1,t_0}, t \ge t_0\}$ is a Cauchy sequence in the L_2 norm and converges in mean square to a random matrix Π_{t_0}. Clearly, for each ω, $\Pi_{t_0}(\omega)$ is a stochastic matrix. Since $\mathbf{1}_{m_1}(\eta_1, \ldots, \eta_{m_1})\Psi_{t_1,t_0}$ in (A.6) is a matrix of identical rows, the mean square error between any two rows of Ψ_{t+1,t_0} is at most $4m_1^2\epsilon$ for all $t \ge t_2$, which implies that Π_{t_0} has identical rows since $\epsilon > 0$ is arbitrary.

We proceed to show that the mean square convergence of Ψ_{t+1,t_0} implies convergence w.p.1. Since $\lim_{t\to\infty} E|\Psi_{t+1,t_0} - \Pi_{t_0}|^2 = 0$, there exists a sequence of integers $t_0 < \tau_1 < \tau_2 < \ldots$ such that $\{\Psi_{\tau_k,t_0}, k \ge 1\}$ converges to Π_{t_0} for all $\omega \in \Omega \backslash N$, where N is a null set.

For any $\omega \in \Omega \setminus N$ and $\epsilon > 0$, there exists k_0 depending on ω such that for all $k \ge k_0$

$$|\Psi_{\tau_k,t_0}(\omega) - \Pi_{t_0}(\omega)| \le \epsilon.$$

For any $t \geq \tau_{k_0}$, since $\Psi_{t,\tau_{k_0}}$ is a stochastic matrix and Π_{t_0} has identical rows

$$\begin{aligned}|\Psi_{t,t_0} - \Pi_{t_0}| &= \left|\Psi_{t,\tau_{k_0}}\Psi_{\tau_{k_0},t_0} - \Pi_{t_0}\right| \\ &= \left|\Psi_{t,\tau_{k_0}}\Psi_{\tau_{k_0},t_0} - \Psi_{t,\tau_{k_0}}\Pi_{t_0}\right| \\ &\leq \left|\Psi_{t,\tau_{k_0}}\right|\left|\Psi_{\tau_{k_0},t_0} - \Pi_{t_0}\right|.\end{aligned}$$

Hence, for each $\omega \in \Omega \setminus N$ and $t \geq \tau_{k_0}$, $|\Psi_{t,t_0}(\omega) - \Pi_{t_0}(\omega)| \leq \sqrt{m_1}\epsilon$, which implies that Ψ_{t,t_0} converges to Π_{t_0} w.p.1. This completes the proof of necessity. □

Appendix B
Proof of Theorem 6

Let $\{e_1, \ldots, e_k\}$ denote the canonical basis of $\mathbb{R}^k$.

Lemma B.1: Let $\{\tilde{M}_t, t \geq 0\}$ be a sequence of $k \times k$ stochastic matrices. Then it has ergodic backward products if and only if

$$Y_{t+1} = \tilde{M}_t Y_t \tag{B.1}$$

achieves consensus (i.e., $\lim_{t\to\infty} Y_t = y^*\mathbf{1}_k$ for some $y^* \in \mathbb{R}$) for any given initial pair (t_0, Y_{t_0}) where $Y_{t_0} \in \{e_1, \ldots e_k\}$, $t_0 \geq 0$.

Proof: Necessity is obvious. We show sufficiency. For any $s \geq 0$, we run the linear system (B.1) by taking the initial pairs (s, e_j), $1 \leq j \leq k$, respectively. Then by consensus

$$\begin{aligned}&\tilde{M}_t \ldots \tilde{M}_{s+1}\tilde{M}_s = \\ &\qquad \tilde{M}_t \ldots \tilde{M}_{s+1}\tilde{M}_s(e_1, \ldots, e_k) \overset{t\to\infty}{\longrightarrow} (c_1\mathbf{1}_k, \ldots, c_k\mathbf{1}_k)\end{aligned}$$

for some constants $c_1, \ldots, c_k$ depending on s. So $\{\tilde{M}_t, t \geq 0\}$ has ergodic backward products. □

Lemma B.2: Let $\{\tilde{M}_t, t \geq 0\}$ be $k \times k$ stochastic matrices with ergodic backward products, and

$$Y_{t+1} = \tilde{M}_t Y_t + \xi_t, \qquad t \geq 0$$

where $\xi_t \in \mathbb{R}^k$ and $\sum_{t=0}^{\infty} |\xi_t| < \infty$. Then $\lim_{t\to\infty} Y_t = y^*\mathbf{1}_k$ for some $y^* \in \mathbb{R}$.

Proof: Denote $\Phi_{t,s} = \tilde{M}_{t-1} \ldots \tilde{M}_s$ for $t > s$, and $\Phi_{s,s} := I$. For $t > t_0 \geq 0$

$$Y_t = \Phi_{t,t_0} Y_{t_0} + \sum_{s=t_0}^{t-1} \Phi_{t,s+1}\xi_s. \tag{B.2}$$

Since $|\Phi_{s_2,s_1}| \leq \sqrt{k}$ for $s_2 \geq s_1$, there exists a fixed C such that $\sup_{t\geq 0} |Y_t| \leq C$.

Given any $\varepsilon > 0$, we may find a sufficiently large t_0 such that

$$\sup_{t_0'' > t_0' \geq t_0} \left| \sum_{s=t_0'}^{t_0''-1} \Phi_{t_0'',s+1}\xi_s \right| \leq \varepsilon. \tag{B.3}$$

Denote $\Phi_{t,t_0}Y_{t_0} = [\eta_{t,1}, \ldots, \eta_{t,k}]^T$ and $Y_t = [Y_t^1, \ldots, Y_t^k]^T$. For a sufficiently large $t_1 > t_0$, by ergodicity we have $\max_{1\leq i,j\leq k} |\eta_{t,i} - \eta_{t,j}| \leq \varepsilon$ for all $t \geq t_1$. Hence, $\max_{1\leq i,j\leq k} |Y_t^i - Y_t^j| \leq 3\varepsilon$ for all $t \geq t_1$. Since ε is arbitrary

$$\lim_{t\to\infty} \max_{1\leq i,j\leq k} \left|Y_t^i - Y_t^j\right| = 0. \tag{B.4}$$

Next, for $t' > t \geq t_1$, we have

$$\begin{aligned}Y_{t'} =& \Phi_{t',t}Y_t + \sum_{s=t}^{t'-1} \Phi_{t',s+1}\xi_s \\ =& Y_t + \left(Y_t^1\mathbf{1}_k - Y_t\right) + \Phi_{t',t}\left(Y_t - Y_t^1\mathbf{1}_k\right) + \sum_{s=t}^{t'-1} \Phi_{t',s+1}\xi_s.\end{aligned}$$

By (B.3), (B.4), $\{Y_t, t \geq 0\}$ is a Cauchy sequence in $\mathbb{R}^k$. By using (B.4) again, there exists y^* such that $\lim_{t\to\infty} Y_t = y^*\mathbf{1}_k$. □

Lemma B.3: Consider (23) with the initial pair $(t_0, \mathbf{X}_{t_0})$. Assume that $\{\mathbf{A}_t, t \geq 0\}$ is (δ_t)-compatible with $\{\tilde{G}_t, t \geq 0\}$ such that (18), (19) hold for $t \geq t_c$. If $d^* \geq 1$, we have

$$\max_{1\leq d\leq d^*} \left|X_t - X_t^{(-d)}\right| \leq C\delta_t^*, \quad t \geq \max\{t_0, t_c\} + d^*$$

where $\delta_t^* = \max_{s\geq 0, t-d^*\leq s\leq t} \delta_s$.

Proof: $\sup_{t\geq t_0} |\mathbf{X}_t| \leq C$ for some constant C. For $t \geq \max\{t_0, t_c\} + d^*$ and any $1 \leq d \leq d^*$, we have $X_t^{(-d)} = X_{t-d}$ and $|A_{00,t-1} - I| \leq C\delta_{t-1}$. Here C may take different values. Hence,

$$\begin{aligned}|X_t - X_{t-1}| &= \left|A_{00,t-1}X_{t-1} + \sum_{d=1}^{d^*} A_{0d,t-1}X_{t-1}^{(-d)} - X_{t-1}\right| \\ &\leq C\delta_{t-1}.\end{aligned} \tag{B.5}$$

Similarly,

$$|X_{t-l+1} - X_{t-l}| \leq C\delta_{t-l}, \quad 1 \leq l \leq d^*.$$

Hence, for any $d \leq d^*$, $|X_t - X_t^{(-d)}| = |X_t - X_{t-d}| \leq C(\delta_{t-1} + \ldots + \delta_{t-d}) \leq C\delta_t^*$. □

Proof of Theorem 6: Sufficiency—Suppose that $\{M_t^A, t \geq 0\}$ has ergodic backward products. Consider (23) with the initial pair $(t_0, \mathbf{X}_{t_0})$. For $t \geq t_0$

$$\begin{aligned}X_{t+1} &= A_{00,t}X_t + \sum_{d=1}^{d^*} A_{0d,t}X_t^{(-d)} \\ &= M_t^A X_t + \sum_{d=1}^{d^*} A_{0d,t}\left(X_t^{(-d)} - X_t\right).\end{aligned}$$

Denote $\xi_t = \sum_{d=1}^{d^*} A_{0d,t}(X_t^{(-d)} - X_t)$. Then by Lemma B.3, $|\xi_t| \leq C\delta_t\delta_t^* \leq C(\delta_t^*)^2$ for $t \geq \max\{t_0, t_c\} + d^*$. So $\sum_{t=t_0}^{\infty} |\xi_t| < \infty$. By Lemma B.2, $\lim_{t\to\infty} X_t = x_\infty\mathbf{1}_n$ for some $x_\infty \in \mathbb{R}$. Hence, $\lim_{t\to\infty} \mathbf{X}_t = x_\infty\mathbf{1}_{n(d^*+1)}$. By Lemma B.1, $\{\mathbf{A}_t, t \geq 0\}$ has ergodic backward products.

Necessity—Suppose that $\{\mathbf{A}_t, t \geq 0\}$ has ergodic backward products. By Proposition 5, $\{M_t^A, t \geq 0\}$ satisfies the (δ_t)-compatibility condition with some constants $t_c, \underline{c}, \bar{c}$. Consider $X_{t+1} = M_t^A X_t$ with the initial pair (t_0, X_{t_0}). Since (20) holds

for M_t^A, $|X_{t+1}-X_t|\le C\delta_t$ for $t\ge\max\{t_0,t_c\}$ and some C. Define $X_{t+1}^{(-1)}=X_t$, $X_{t+1}^{(-d)}=X_t^{(-(d-1))}$ for $d=2,\dots,d^*$. Fix any initial condition $(X_{t_0},X_{t_0}^{(-1)},\dots,X_{t_0}^{(-d^*)})$. By using the method in proving Lemma B.3, we may show that

$$\max_{1\le d\le d^*}\left|X_t-X_t^{(-d)}\right|\le C\delta_t^*,\quad t\ge\max\{t_0,t_c\}+d^*.$$

Then

$$X_{t+1}=A_{00,t}X_t+\sum_{d=1}^{d^*}A_{0d,t}X_t^{(-d)}+\sum_{d=1}^{d^*}A_{0d,t}\left(X_t-X_t^{(-d)}\right).$$

Letting $\mathbf{X}_t=[X_t;X_t^{(-1)};\dots;X_t^{(-d^*)}]$, we may write

$$\mathbf{X}_{t+1}=\mathbf{A}_t\mathbf{X}_t+\xi_t,\quad t\ge t_0$$

where $\xi_t=[\sum_{d=1}^{d^*}A_{0d,t}(X_t-X_t^{(-d)});0_{nd^*\times1}]$. We have $\sum_{t=t_0}^{\infty}|\xi_t|<\infty$. Since $\{\mathbf{A}_t,t\ge0\}$ has ergodic backward products, by Lemma B.2 there exists x_∞ such that $\lim_{t\to\infty}\mathbf{X}_t=x_\infty\mathbf{1}_{n(d^*+1)}$, which implies $\lim_{t\to\infty}X_t=x_\infty\mathbf{1}_n$. Since (t_0,X_{t_0}) is arbitrary, Lemma B.1 implies that $\{M_t^A,t\ge0\}$ has ergodic backward products. □

Appendix C
Proof of Theorem 8

Lemma C.1: Let $t_1\ge2$, $T\ge0$, and $\gamma\in(0,1]$. Then $\sum_{t=t_1}^{t_1+T}t^{-\gamma}\le(T+1)/(t_1-1)^\gamma$.

Proof: If $\gamma\in(0,1)$, we have

$$\begin{aligned}\sum_{t=t_1}^{t_1+T}t^{-\gamma}&\le\int_{t_1-1}^{t_1+T}t^{-\gamma}dt\\&=\frac{(t_1-1)^{1-\gamma}}{1-\gamma}\left\{[1+(T+1)/(t_1-1)]^{1-\gamma}-1\right\}\\&\le\frac{(t_1-1)^{1-\gamma}}{1-\gamma}(1-\gamma)\frac{T+1}{t_1-1}=\frac{T+1}{(t_1-1)^\gamma}.\end{aligned}$$

If $\gamma=1$, the estimate is similar and the detail is omitted. □

Proof of Theorem 8: We say that condition (C1) is satisfied infinitely often (i.o.) by elements in a sequence $\{h_t,t\ge T_0\}$ if given any $T_1\ge T_0$, there exists $t\ge T_1$ such that h_t satisfies (C1). It suffices to show $\lim_{t\to\infty}Z_t=c\mathbf{1}_n$. We consider the initial pair $(0,X_0)$. The proof with a general initial pair (t_0,X_{t_0}) is similar. We prove by induction for the n components of Z_t.

Step 1) By Proposition 7, $\lim_{t\to\infty}z_t^1=z_\infty^1$ for some finite z_∞^1.

Step 2) Assume that for $l\in\{1,\dots,n-1\}$,

$$\lim_{t\to\infty}z_t^k=z_\infty^1\tag{C.1}$$

for all $k\le l$. Next we show that (C.1) holds for $k=l+1$, which is given in several sub-steps.

Step 2.1 (Contradiction argument) Suppose that (C.1) is not true for $k=l+1$, which implies that there exists $0<\epsilon_0\le1$ such that

$$\left|z_t^{l+1}-z_\infty^1\right|\ge\epsilon_0\quad\text{i.o.}\tag{C.2}$$

Since $\limsup_{t\to\infty}z_t^{l+1}\le\lim_{t\to\infty}z_t^1=z_\infty^1$ by the ordering of the elements in Z_t, (C.2) implies that

$$z_t^{l+1}\le z_\infty^1-\epsilon_0\quad\text{i.o.}\tag{C.3}$$

Step 2.2 (Estimate of the $(n-l)$ lowest trajectories) By compatibility, suppose that (20), (21) hold with $\delta_t=a_t$, $t\ge t_c\ge0$. For any $\epsilon_1\in(0,\epsilon_0/3)$, by the induction assumption (C.1), there exists $T(\epsilon_1)>t_c$ such that for all $t\ge T(\epsilon_1)$

$$\left|z_t^k-z_\infty^1\right|\le\epsilon_1,\quad k=1,\dots,l.\tag{C.4}$$

By (C.3), there exists $t_1\ge\max\{T(\epsilon_1),2\}$ such that $z_{t_1}^j\le z_\infty^1-\epsilon_0$ for $j=l+1,\dots,n$ since $z_{t_1}^{l+1}\ge z_{t_1}^{l+2}\ge\dots\ge z_{t_1}^n$. Consider the vector Z_{t_1} and use our convention of associating a component with a node. Suppose

$$z_{t_1}^1=x_{t_1}^{i_1},\ z_{t_1}^2=x_{t_1}^{i_2},\ \dots,\ z_{t_1}^l=x_{t_1}^{i_l}\tag{C.5}$$

for $l\ge1$ distinct elements $i_1,\dots,i_l$ in $\mathcal{N}=\{1,\dots,n\}$, which implicitly depend on t_1. Denote $S_{t_1}=\{i_1,\dots,i_l\}$ and its complement $S_{t_1}^c=\mathcal{N}\setminus S_{t_1}$. Note that $\sup_{t\ge0,1\le k\le n}|x_t^k|\le C_{X_0}:=\max_{1\le i\le n}|x_0^i|$. Denote $M_t=(a_{ij}(t))_{1\le i,j\le n}$. For $j\in S_{t_1}^c$ and $t\ge t_1$, it follows from (20) with the substitution of $\delta_t=a_t$ that

$$\begin{aligned}x_{t+1}^j&=\sum_{k=1}^n a_{jk}(t)x_t^k\\&=\sum_{k\in S_{t_1}^c}a_{jk}(t)x_t^k+\sum_{k\in S_{t_1}}a_{jk}(t)x_t^k\\&\le\max_{k\in S_{t_1}^c}x_t^k+c_1/t^\gamma\end{aligned}\tag{C.6}$$

where c_1 depends only on $\{M_t,t\ge0\}$ and C_{X_0}. Since $j\in S_{t_1}^c$ is arbitrary, (C.6) implies that

$$\max_{k\in S_{t_1}^c}x_{t+1}^k\le\max_{k\in S_{t_1}^c}x_t^k+c_1/t^\gamma,\quad t\ge t_1.\tag{C.7}$$

Subsequently, we choose a large T such that

$$\sum_{t=t_1}^{t_1+T}c_1/t^\gamma\le\epsilon_0/3.\tag{C.8}$$

By Lemma C.1, we take

$$T=\lfloor\epsilon_0(t_1-1)^\gamma/(3c_1)-1\rfloor\tag{C.9}$$

to ensure (C.8). We use $\lfloor r\rfloor$ to denote the greatest integer bounded from above by r. Without loss of generality, assume that given

(ϵ_0, ϵ_1), a sufficiently large t_1 has been selected such that

$$\frac{\epsilon_0(t_1-1)^\gamma}{3c_1} \geq 4. \tag{C.10}$$

The iteration of (C.7) yields

$$\begin{aligned}\max_{t_1\leq t\leq t_1+T+1}\max_{k\in S_{t_1}^c} x_t^k \leq & \max_{k\in s_{t_1}^c} x_{t_1}^k + \\ & c_1\sum_{t=t_1}^{t_1+T} 1/t^\gamma \\ & \leq z_{t_1}^{l+1} + \epsilon_0/3 \leq z_\infty^1 - \epsilon_0 + \epsilon_0/3 \\ & = z_\infty^1 - (2\epsilon_0)/3. \end{aligned} \tag{C.11}$$

By (C.4) and (C.11), the states of nodes $i_1, \ldots, i_l$ remain to be the first l greatest states for $t_1 \leq t \leq t_1+T+1$ although the position of these states may change with time when listed in Z_t. Now by $\epsilon_1 < \epsilon_0/3$, it follows from (C.4) and (C.11) that for any $i \in S_{t_1}$ and any $j \in S_{t_1}^c$

$$\left|x_t^i - z_\infty^1\right| \leq \epsilon_1, \quad x_t^j \leq x_t^i - \epsilon_0/3 \tag{C.12}$$

where $t_1 \leq t \leq t_1+T+1$ and T is defined by (C.9).

Step 2.3 (Estimate of the l highest trajectories) For node $i \in S_{t_1}$

$$\begin{aligned} x_{t+1}^i = & \sum_{k\in S_{t_1}} a_{ik}(t)x_t^k + \sum_{k\in S_{t_1}^c} a_{ik}(t)x_t^k \\ = & x_t^i + [a_{ii}(t)-1]\,x_t^i + \sum_{k\in S_{t_1}\setminus\{i\}} a_{ik}(t)x_t^k \\ & + \sum_{k\in S_{t_1}^c} a_{ik}(t)x_t^k \\ = & x_t^i + \sum_{k\in S_{t_1}\setminus\{i\}} a_{ik}(t)\left(x_t^k - x_t^i\right) \\ & + \sum_{k\in S_{t_1}^c} a_{ik}(t)\left(x_t^k - x_t^i\right) \end{aligned} \tag{C.13}$$

where $t_1 \leq t \leq t_1+T$. It follows from (C.12) that

$$\begin{aligned}\left|\sum_{k\in S_{t_1}\setminus\{i\}} a_{ik}(t)\left(x_t^k - x_t^i\right)\right| \leq & 2\epsilon_1 \sum_{k\in S_{t_1}\setminus\{i\}} a_{ik}(t) \\ \leq & (2\epsilon_1)c_2 t^{-\gamma}\end{aligned}$$

where c_2 depends only on $\{M(t), t \geq 0\}$. We have

$$\sum_{k\in S_{t_1}^c} a_{ik}(t)\left(x_t^k - x_t^i\right) \leq \sum_{k\in S_{t_1}^c} a_{ik}(t)(-\epsilon_0/3).$$

Denote $g_t = \sum_{i\in S_{t_1}} x_t^i$. Hence, by (C.13),

$$g_{t+1} \leq g_t + 2\epsilon_1 c_2 n t^{-\gamma} - \left(\frac{\epsilon_0}{3}\right)\sum_{i\in S_{t_1}}\sum_{k\in S_{t_1}^c} a_{ik}(t). \tag{C.14}$$

Since $\tilde{G}_t$ is strongly connected, there exists a pair $(i,k) \in S_{t_1} \times S_{t_1}^c$ such that (k,i) is an edge of $\tilde{G}_t$ and $a_{ik}(t) \geq c_3 t^{-\gamma}$ for $t \geq t_c$ by compatibility, where $c_3 > 0$ is determined by $\{M_t, t \geq 0\}$.

Subsequently, it follows that

$$g_{t+1} \leq g_t + 2\epsilon_1 c_2 n t^{-\gamma} - \left(\frac{\epsilon_0}{3}\right) c_3 t^{-\gamma} \tag{C.15}$$

where $t_1 \leq t \leq t_1+T$. Since c_2 and c_3 depend only on $\{M(t), t \geq 0\}$, we may further assume that a sufficient small ϵ_1 has been selected such that

$$2\epsilon_1 c_2 n < \frac{\epsilon_0 c_3}{6}.$$

Iterating (C.15), we have

$$g_{t_1+T+1} \leq g_{t_1} - \sum_{t=t_1}^{t_1+T}\left(\frac{\epsilon_0 c_3}{6}\right) t^{-\gamma}.$$

Step 2.4 (A contradiction) Since $t_1 \geq 2$ and (C.10) holds, for T given by (C.9), it follows that

$$\begin{aligned} T \geq & \epsilon_0(t_1-1)^\gamma/(3c_1) - 2 \\ & \geq \epsilon_0(t_1-1)^\gamma/(6c_1) \geq \epsilon_0 t_1^\gamma/(12c_1). \end{aligned}$$

Denote $\eta = \epsilon_0/(12c_1)$. Then $\eta \leq 1/(12c_1)$ since $\epsilon_0 \leq 1$.

(a) If $0 < \gamma < 1$

$$\begin{aligned}\sum_{t=t_1}^{t_1+T} t^{-\gamma} \geq & \int_{t_1}^{t_1+T} v^{-\gamma}dv \geq \int_{t_1}^{t_1+\eta t_1^\gamma} v^{-\gamma}dv \\ & = t_1^{1-\gamma}\left[\left(1+\eta t_1^{\gamma-1}\right)^{1-\gamma} - 1\right]/(1-\gamma).\end{aligned}$$

For $f(v) = (1+v)^{1-\gamma}$, $v \in [0, 1/(12c_1)]$, there exists a constant $D_1 > 0$ depending only on γ and $c_1 > 0$ such that $f(v) \geq 1 + D_1 v$ for all $v \in [0, 1/(12c_1)]$. Now $\eta t_1^{\gamma-1} \leq 1/(12c_1)$ and

$$\begin{aligned}\sum_{t=t_1}^{t_1+T} t^{-\gamma} \geq & t_1^{1-\gamma}\left(1 + D_1\eta t_1^{\gamma-1} - 1\right)/(1-\gamma) \\ & = D_1\epsilon_0/\left[12c_1(1-\gamma)\right].\end{aligned}$$

So for $0 < \gamma < 1$, we obtain

$$g_{t_1+T+1} \leq g_{t_1} - D_1\epsilon_0^2 c_3/\left[72c_1(1-\gamma)\right]. \tag{C.16}$$

On the other hand, by (C.12) we have

$$|g_{t_1+T+1} - g_{t_1}| \leq 2n\epsilon_1. \tag{C.17}$$

If we take

$$0 < \epsilon_1 < \min\left\{\frac{\epsilon_0}{3}, \epsilon_0 c_3/(12nc_2), D_1\epsilon_0^2 c_3/\left[144nc_1(1-\gamma)\right]\right\}$$

then (C.16), (C.17) lead to a contradiction. Hence, (C.1) holds for $k = l+1$.

(b) If $\gamma = 1$, a similar argument may be used to show (C.1) for $k = l+1$ and the detail is omitted.

Step 3) Finally, we conclude that (C.1) holds for all $k \in \mathcal{N}$. This completes the proof. □

APPENDIX D

Lemma D.1: Assume condition i) in Theorem 11 holds. Let $h \geq 1$ be a fixed integer and denote $\tilde{G}_{[t,t+s)} = (\mathcal{N}, \tilde{\mathcal{E}}_{[t,t+s)})$ for $t \geq 0$ and $1 \leq s \leq h$. Then there exist constants $t_c' \geq 1$ and $0 < \underline{c}' \leq \overline{c}'$, all independent of (t,s), such that for all $t \geq t_c'$, $M^{t,s} := M_{t+s-1} \dots M_t$ satisfies

$$M^{t,s}(i,j) \leq \overline{c}' t^{-\gamma}, \quad \forall 1 \leq i,j \leq n, \quad j \neq i \tag{D.1}$$

$$M^{t,s}(i,j) \geq \underline{c}' t^{-\gamma}, \quad \forall (j,i) \in \tilde{\mathcal{E}}_{[t,t+s)} \tag{D.2}$$

where $M^{t,s}(i,j)$ is the (i,j)th element of $M^{t,s}$.

Proof: If $s = 1$, the lemma is obvious. Below we consider $s \geq 2$ and $t \geq t_c$, where t_c is selected such that (20), (21) hold for $t \geq t_c$ with constants t_c, $0 < \underline{c} \leq \overline{c}$.

For $j \neq i$, we have

$$M^{t,s}(i,j) = \sum_{k_1,\dots,k_{s-1}} M_{t+s-1}(i,k_{s-1}) \dots M_{t+1}(k_2,k_1) M_t(k_1,j)$$

where each $k_l \in \{1,\dots,n\}$. Each factor of $M_{t+s-1}(i,k_{s-1}) \dots M_{t+1}(k_2,k_1)M_t(k_1,j)$ is between 0 and 1. Since $j \neq i$, there is at least one factor of the form $M_{t+r}(i,k)$ with $k \neq i$ and $0 \leq r \leq s-1$, so that it is bounded from above by $\overline{c}a_{t+r}$ by condition i). In addition, the number of terms in the summation has a finite upper bound depending only on h. Hence (D.1) follows.

We proceed to prove (D.2). Assume that $(j,i) \in \tilde{G}_{[t,t+s)}$. So there exists $0 \leq \xi \leq s-1$ such that $(j,i) \in \tilde{G}_{t+\xi}$. Consequently, by the compatibility condition $M_{t+\xi}(i,j) \geq \underline{c}a_{t+\xi}$. We have

$$\begin{aligned} M^{t,s}(i,j) &= \sum_{k_1,\dots,k_{s-1}} M_{t+s-1}(i,k_{s-1}) \dots M_{t+1}(k_2,k_1) M_t(k_1,j) \\ &\geq M_{t+s-1}(i,i) \dots M_{t+\xi+1}(i,i) M_{t+\xi}(i,j) M_{t+\xi-1}(j,j) \\ &\quad \dots M_t(j,j). \end{aligned} \tag{D.3}$$

If $\xi = 0$ (resp., $\xi = s-1$), $M_{t+\xi}(i,j)$ takes the position of the most right (resp., left) term in (D.3).

By (20), there exist $t_c' \geq t_c$ and $\delta > 0$ such that $\min_{1\leq i\leq n} M_{t+r}(i,i) \geq 1 - \delta a_{t+r} \geq 1/2$ for $t \geq t_c'$ and $r \geq 0$. We obtain

$$\begin{aligned} M^{t,s}(i,j) \geq (1-\delta a_{t+s-1}) \dots (1-\delta a_{t+\xi+1}) \underline{c} a_{t+\xi} \\ \times (1-\delta a_{t+\xi-1}) \dots (1-\delta a_t). \end{aligned}$$

Since $0 \leq \xi < s \leq h$ for the fixed h, by (**A1**) we may select t_c' such that (D.2) holds. □

Lemma D.2: Assume conditions i)–ii) in Theorem 11 holds. Then $\{\widehat{M}_t, t \geq 0\}$ is (a_t)-compatible with $\{\widehat{G}_t, t \geq 0\}$, where $\widehat{M}_t = M_{\tau_{t+1}-1} \dots M_{\tau_t}$ and $\widehat{G}_t = \tilde{G}_{[\tau_t,\tau_{t+1})}$.

Proof: There exists a fixed constant $h > 0$ such that

$$1 \leq \tau_{t+1} - \tau_t \leq h, \quad t \leq \tau_t \leq ht, \quad t \geq 0.$$

So the lemma follows from Lemma D.1. □

ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers and the Associate Editor Prof. C. Szepesvári for very useful comments and suggestions which have helped improve the presentation of the paper.

REFERENCES

[1] J. Abounadi, D. P. Bertsekas, and V. Borkar, "Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms," *SIAM J. Control Optim.*, vol. 41, no. 1, pp. 1–22, 2002.

[2] D. Acemoglu, A. Nedić, and A. Ozdaglar, "Convergence of rule-of-thumb learning rules in social networks," in *Proc. 47th IEEE CDC*, Cancun, Mexico, Dec. 2008, pp. 1714–1720.

[3] T. C. Aysal and K. E. Barner, "Convergence of consensus models with stochastic disturbances," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 4101–4113, Aug. 2010.

[4] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, Oct. 2008.

[5] F. Bćnćzit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *Proc. Int. Symp. Inf. Theory*, Austin, TX, Jun. 2010, pp. 1753–1757.

[6] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," in *Proc. 44th IEEE CDC-ECC'05*, Seville, Spain, Dec. 2005, pp. 2996–3001.

[7] V. S. Bokar, "Asynchronous stochastic approximation," *SIAM J. Control Optim.*, vol. 36, no. 3, pp. 840–851, 1998, Erratum: vol. 38, no. 2, pp. 662–663, 2000.

[8] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[9] P. Brćmaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York: Springer-Verlag, 1999.

[10] R. Carli, F. Fagnani, P. Frasca, T. Taylor, and S. Zampieri, "Average consensus on networks with transmission noise or quantization," in *Proc. Eur. Control Conf.*, Kos, Greece, Jul. 2007, pp. 1852–1857.

[11] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, "Communication constraints in the average consensus problem," *Automatica*, vol. 44, pp. 671–684, 2008.

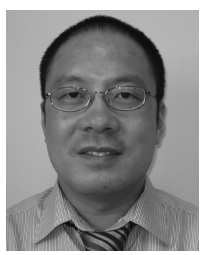
[12] F. Cucker and E. Mordecki, "Flocking in noisy environments," *J. Math. Pures Appl.*, vol. 89, no. 3, pp. 278–296, 2008.

[13] L. Fang and P. J. Antsaklis, "Asynchronous consensus protocols using nonlinear paracontraction theory," *IEEE Trans. Autom. Control*, vol. 53, no. 10, pp. 2351–2355, Nov. 2008.

[14] R. Gharavi and V. Anantharam, "Structure theorems for partially asynchronous iterations of a nonnegative matrix with random delays," *Sādhanā*, vol. 24, no. 4–5, pp. 369–423, 1999.

[15] B. Gharesifard and J. Cortés, "Distributed strategies for generating weight-balanced and doubly stochastic digraphs," *Eur. J. Control*, 2012, to be published.

[16] Y. Hatano and M. Mesbahi, "Agreement in random networks," *IEEE Trans. Autom. Control*, vol. 51, no. 11, pp. 1867–1872, Nov. 2005.
[17] M. Huang, S. Dey, G. N. Nair, and J. H. Manton, "Stochastic consensus over noisy networks with Markovian and arbitrary switches," *Automatica*, vol. 46, no. 10, pp. 1571–1583, 2010.
[18] M. Huang and J. H. Manton, "Coordination and consensus of networked agents with noisy measurements: Stochastic algorithms and asymptotic behavior," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 134–161, 2009.
[19] M. Huang and J. H. Manton, "Stochastic consensus seeking with noisy and directed inter-agent communication: Fixed and randomly varying topologies," *IEEE Trans. Autom. Control*, vol. 55, no. 1, pp. 235–241, Jan. 2010.
[20] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 988–1000, Jun. 2003.
[21] 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.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.
[22] A. Kashyap, T. Başar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.
[23] V. Krishnamurthy, K. Topley, and G. Yin, "Consensus formation in a two-time-scale Markovian system," *SIAM J. Multiscale Model. Simul.*, vol. 7, no. 4, pp. 1898–1927, 2009.
[24] H. J. Kushner and G. Yin, "Stochastic approximation algorithms for parallel and distributed processing," *Stochastics*, vol. 22, no. 3, pp. 219–250, 1987.
[25] I. K. L. Elsner and M. Neumann, "On the convergence of asynchronous paracontractions with applications to tomographic reconstruction from incomplete data," *Linear Algebra and Appl.*, vol. 130, pp. 65–82, 1990.
[26] I. K. L. Elsner and M. Neumann, "Convergence of sequential and asynchronous nonlinear paracontractions," *Numer. Math.*, vol. 62, pp. 305–319, 1992.
[27] T. Li and J.-F. Zhang, "Mean square average-consensus under measurement noises and fixed topologies," *Automatica*, vol. 45, no. 8, pp. 1929–1936, 2009.
[28] T. Li and J.-F. Zhang, "Consensus conditions of multi-agent systems with time-varying topologies and stochastic communication noises," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2043–2057, Sep. 2010.
[29] I. Matei, N. Martins, and J. S. Baras, "Almost sure convergence to consensus in Markovian random graphs," in *Proc. 47th IEEE CDC*, Cancun, Mexico, Dec. 2008, pp. 3535–3540.
[30] M. Nourian, P. E. Caines, R. P. Malhamć, and M. Huang, "Derivation of consensus algorithm dynamics from mean field stochastic control NCE equations," in *Proc. 1st IFAC Workshop Estimation and Control of Netw. Syst.*, Venice, Italy, Sep. 2009.
[31] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.
[32] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1520–1533, Sep. 2004.
[33] A. Olshevsky and J. N. Tsitsiklis, "On the nonexistence of quadratic Lyapunov functions for consensus algorithms," *IEEE Trans. Autom. Control*, vol. 53, no. 11, pp. 2642–2645, Dec. 2008.
[34] S. Patterson, B. Bamieh, and A. El Abbadi, "Convergence rates of distributed average consensus with stochastic link failures," *IEEE Trans. Autom. Control*, vol. 55, no. 4, pp. 880–892, Apr. 2010.
[35] R. Rajagopal and M. J. Wainwright, Network-based consensus averaging with general noisy channels, 2008 [Online]. Available: http://arxiv.org/abs/0805.0438
[36] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Trans. Autom. Control*, vol. 50, no. 5, pp. 655–661, May 2005.
[37] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proc. Amer. Control Conf.*, Portland, OR, Jun. 2005, pp. 1859–1864.
[38] W. Ren, R. W. Beard, and D. B. Kingston, "Multi-agent Kalman consensus with relative uncertainty," in *Proc. Amer. Control Conf.*, Portland, OR, Jun. 2005, pp. 1865–1870.
[39] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links-part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.
[40] E. Seneta, *Non-negative Matrices and Markov Chains*, 2nd ed. New York: Springer, 2006, (revised printing).
[41] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," in *Proc. 46th IEEE CDC*, New Orleans, LA, Dec. 2007, pp. 1535–1540.
[42] B. Touri and A. Nedić, "Distributed consensus over network with noisy links," in *Proc. 12th Int. Conf. Info. Fusion*, Seattle, WA, Jul. 2009, pp. 146–154.
[43] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learn.*, vol. 16, pp. 185–202, 1994.
[44] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.
[45] J. Wolfowitz, "Products of indecomposable, aperiodic, stochastic matrices," *Proc. American Math. Soc.*, vol. 14, no. 5, pp. 733–737, 1969.
[46] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, pp. 33–46, 2007.
[47] G. F. Young, L. Scardovi, and N. E. Leonard, "Robustness of noisy consensus dynamics with directed communication," in *Proc. Amer. Control Conf.*, Baltimore, MD, Jun. 2010, pp. 6312–6317.



Minyi Huang (M'03) received the B.Sc. degree from Shandong University, Jinan, Shandong, China, in 1995, the M.Sc. degree from the Institute of Systems Science, Chinese Academy of Sciences, Beijing, China, in 1998, and the Ph.D. degree from the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, in 2003, all in the area of systems and control.

He was a Research Fellow first in the Department of Electrical and Electronic Engineering, the University of Melbourne, Australia, from February 2004 to March 2006, and then in the Department of Information Engineering, Research School of Information Sciences and Engineering, the Australian National University, Canberra, Australia, from April 2006 to June 2007. He joined the School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada, as an Assistant Professor in July 2007, where he has been an Associate Professor since July 2011. His research interests include mean field stochastic control and dynamic games, multi-agent control and computation in distributed networks with applications.