CONSENSUS-TYPE STOCHASTIC APPROXIMATION ALGORITHMS

by

YU SUN

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2012

MAJOR: MATHEMATICS

Approved by:

____
Advisor Date

____

____

____

DEDICATION

To My Parents

Chunrong Fang and Haiqing Sun

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the contributions of many individuals. In particular, I wish to express my deepest gratitude and sincerest appreciation to my advisor Professor George Yin for proposing me an exciting topic for my dissertation, for his exceptional guidance, constant inspiration, consistent support, and endless care during my graduate study at Wayne State University. His direction and support were crucial for the completion of the work.

I would also like to thank Professor Le Yi Wang for introducing me to the area of consensus problems. I greatly admire his engineering perspectives, and I am thankful for the opportunity to have learned from him. His comments, suggestions and suppot have resulted in much improvement of this work.

I am taking this opportunity to thank Professor Rafail Z. Khasminskii, Professor Tze-Chien Sun, and Professor Fatih Celiker for serving on my committee.

I am grateful to all my professors, including the Chair, Professor Daniel Frohardt, Professor Boris Mordukhovich, Professor Dan Isaksen, Professor Alex Korostelev, and Professor Guozhen Lu, who have contributed to my solid formation in mathematics, who have nurtured my curiosity and given me the opportunities to learn from them.

My thanks also go to Ms. Mary Klamo, Ms. Patricia Bonesteel, Paul-Liu Chow, Professor John Breckenridge, Professor Choon-Jai Rhee, Dr. Yuanjin Liu, Professor Jinfeng Wei, Professor Shaobai Kan, Guangliang Zhao, Ren Zhao, Qi He and Xiaolong Han for their help and encouragement in one way or another.

I am indebted to my beloved parents for their love, support, and inspiration since

I was born. This dissertation is dedicated to my dearest Mother Chunrong Fang and Father Haiqing Sun.

Special thanks go to my fiance, Dingzhou Cao for his love, support, encouragement, and his confidence in me, especially when I lacked myself. He is my permanent source of love and strength. The Ph.D. journey was colorful and memorable because of his sweet love.

Finally, but not least, I would like to express my appreciation to the entire Department of Mathematics for their hospitality and services. I enjoyed the warm and friendly atmosphere in the department, and I appreciated the support I received during my study at Wayne State University.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

Consensus controls represent a team effort to reach a common goal. The problems are related to many applications that involve coordination of multiple entities with only limited neighborhood information to reach a global goal for the entire team. Typical examples include multi-agents in robotics, flocking behavior in people and animals, wireless communication networks, sensor networks, platoon formation in ground and aerial vehicles, distributed computing, biological systems, etc. Due to the diversity in application domains, detailed system descriptions vary substantially and diversified methodologies are needed to treat such systems. However, one common feature of the underlying problems is: Although the goal of control is global to the entire system, only limited local information is available for control actions.

There is an extensive literature on consensus control in a variety of application areas, including computing load balancing [22, 40], sensor networks [1, 28], mobil agents [13, 27], flocking behavior and swarms [21, 36, 37], etc. Related algorithms and theoretical developments were reported in [5, 11, 31]. Much of recent work was motivated by [37], which in fact is a version of a model introduced earlier in [33] for simulating flocking and schooling behaviors in computer graphics. The effort in the control community can be traced back to the asynchronous stochastic optimization algorithms [38], which was substantially generalized in [18]. In this dissertation, we consider a specific control structure for consensus. It is noted that consensus control often leads to consensus without further constraints on the actual state. Practical systems always require states to be confined in some ways. Our link-based control provides a natural and practical constraint on the state.

With the aforementioned motivations, this dissertation is mainly concerned with the study of convergence properties of consensus-type algorithms for networked systems in which the network topologies switch randomly and develops an iterate averaging algorithm for consensus-type controls of networked systems. Our interest in this problem is motivated by cooperative and coordinated control. Owing to the wide variety of applications, detailed system descriptions vary substantially and diversified methodologies are needed to treat such systems. Nevertheless, there is a common thread, the use of an online recursive stochastic approximation (SA) algorithm. There is extensive literature on consensus control in a variety of application areas, including load balancing in parallel computing [22, 40, 38, 42], sensor networks [1, 9, 28, 29], team formation [4], decentralized filtering, estimation, and data fusion [3, 17, 23, 4, 35], mobile agents [13, 26, 31, 32], flocking behavior and swarms [21, 33, 37], physics [37], etc. Applications of stochastic approximation algorithms and theoretical developments in related consensus control problems were reported in [5, 10, 11]. Switching network topologies were studied in [24, 26, 12]. More recently, [14] employed a method on the convergence of products of stochastic matrices that uses randomly switching Laplacian matrices together with observation noises that may be state-dependent and Markovian based. In [47], we used a Markov model and treated a much larger class of noises, where the network graph is modulated by a discrete-time Markov chain. In addition to convergence and rates of convergence, a multi-scale structure, which captures differences between state adaptation speeds and topology switching frequencies, was explored fully. Related stochastic differential equations and switching stochastic equations were obtained.

As an application, imagine we have a collection of UAV assets tasked with searching a forward operating location for the presence of targets. Decisions must be made about individual UAV task assignments, since the collections of UAVs might be heterogeneous with regard to capabilities. In addition, tracking potential targets over long distances may require"target hand-off" that must be coordinated among teams of UAVs. As another application, we consider for instance the problem of networked computing [40, 48]. A computational job is assigned to a network of r computers. The goal is to achieve approximately equal workload distribution for each computer to avoid idle or overloaded running states. A workload transfer from node i to node j results in a decrease of workload at node i and an increase of the same amount at node j. This control structure does not change the total workload amount of the whole system and provides a natural constraint to bound the node states. This scenario can be easily recognized in different application domains such as material distribution systems, data fusion in distributed sensor networks, deployment of sensors, coordination of unmanned aerial vehicles (UAVs). It will be shown that this constraint leads to a Markovian dynamic system that connects seamlessly with the Markov chain descriptions of the network topology switching dynamics.

To model inherent uncertainties, this dissertation considers consensus control problems with regime-switching network topologies. In our setup, we quantify the time-varying parameter process as a Markov chain with a transition matrix that includes a small parameter $\varepsilon > 0$, which characterizes the rates of network switching. We then use a stochastic recursive algorithm to carry out the consensus control task. The algorithm uses a small stepsize $\mu > 0$, which defines how fast the network node states are updated. The impact of network switching rates on convergence properties of consensus control algorithms is captured by the relationship between ε and μ. There are three cases concerning the relative sizes of ε and μ: $0 < \varepsilon = \mathcal{O}(\mu)$, $0 < \varepsilon \ll \mu$, and $0 < \mu \ll \varepsilon$. Asymptotic behaviors of consensus control algorithms under these cases are fundamentally different. When $\varepsilon = \mathcal{O}(\mu)$, through appropriate interpolations, the limit is described by regime-switching ordinary differential equations. When $\varepsilon \ll \mu$, the network topology rarely changes and is essentially fixed during the transient interval of active consensus control. We thus practically deal with a fixed network. When $\mu \ll \varepsilon$, the network is changing so fast that it acts like a noise, and consequently only its average with respect to the stationary measure determines convergence properties of the consensus control.

To summarize, in this dissertation, we investigate the asymptotic properties of consensus-type algorithms using iterate averaging and regime-switching topologies. In each setting, theoretical results (e.g., algorithms, convergence, and asymptotic efficiency, ect.) are developed, and numerical experiments are presented to illustrate the tracking performance of the identification algorithms.

The remainder of the dissertation is arranged as follows. Chapter 2 introduces the networked systems and consensus control problems. Some basic properties of networked systems are derived for time-invariant systems, which are to be used in subsequent convergence analysis. We propose the two-stage recursive algorithm in Chapter 3. We obtains convergence, rate of convergence, and asymptotic efficiency using stochastic approximation iterate averaging algorithms. Numerical examples to illustrate the asymptotes are provided. Chapter 4 sets the stage for networked systems with randomly time-varying topologies. The problem formulation of regime-switching network topologies is introduced. Convergence analysis under the scenario $\varepsilon = \mathcal{O}(\mu)$ is presented. We also focuses on convergence analysis for the cases of fast-switching and slow-switching network topologies, as well as simulation examples to illustrate the asymptotes. Finally, we end this dissertation with conclusions and further remarks in Chapter 5.

2 Networked System and Consensus Control

Consider a networked system of r nodes, given by

$$x_{n+1}^i = x_n^i + u_n^i, \quad i = 1, \ldots, r, \tag{2.1}$$

where u_n^i is the node control for the ith node, or in a vector form $x_{n+1} = x_n + u_n$ with $x_n = [x_n^1, \ldots, x_n^r]'$, $u_n = [u_n^1, \ldots, u_n^r]'$. The nodes are linked by a sensing network, represented by a directed graph $\mathcal{G}$ whose element (i, j) indicates estimation of the state x_n^j by node i via a communication link, and a permitted control v^{ij} on the link. For node i, $(i, j) \in \mathcal{G}$ is a departing edge and $(l, i) \in \mathcal{G}$ is an entering edge. The total number of communication links in $\mathcal{G}$ is l_s. From its physical meaning, node i can always observe its own state, which will not be considered as a link in $\mathcal{G}$.

2.1 Networked Observation and Control

In this dissertation, we limit the control structures to the link control among nodes permitted by $\mathcal{G}$. The node control u_n^i is determined by the link control v_n^{ij}. Since a positive transportation of quantity v_n^{ij} on (i, j) means a loss of v_n^{ij} at node i and a gain of v_n^{ij} at node j, the node control at node i is $u_n^i = -\sum_{(i,j)\in\mathcal{G}} v_n^{ij} + \sum_{(j,i)\in\mathcal{G}} v_n^{ji}$. The most relevant implication in this control scheme is that for all n, $\sum_{i=1}^r x_n^i = \sum_{i=1}^r x_0^i := \eta r$, for some $\eta \in \mathbb{R}$ that is the average of x_0. That is, $\eta = \sum_{i=1}^r x_0^i / r$. Consensus control seeks control algorithms that achieve $x_n \to \eta \mathbb{1}$, where $\mathbb{1}$ is the column vector of all 1s. A link $(i, j) \in \mathcal{G}$ entails an estimate, denoted by $\widehat{x}_n^{ij}$, of x_n^j by node i with estimation error d_n^{ij}, i.e.,

$$\widehat{x}_n^{ij} = x_n^j + d_n^{ij}. \tag{2.2}$$

The estimation error d_n^{ij} is usually a function of the signal x_n^j itself and depends on communication channel noises ξ_n^{ij} in a nonadditive and nonlinear relation

$$d_n^{ij} = g(x_n^j, \xi_n^{ij}) \tag{2.3}$$

and can be spatially and temporally dependent. Most existing literature considers much simplified noise classes $d_n^{ij} = \xi_n^{ij}$ with i.i.d. assumptions.

This dissertation will consider general noise classes of type (2.3). Such extensions are necessary when dealing with networked systems. A sampled and quantized signal x in a networked system enters a communication transmitter as a source. To enhance channel efficiency and reduce noise effects, source symbols are encoded [6, 15]. Typical block or convolutional coding schemes such as Hamming, Reed-Solomon, or more recently the low-density parity-check (LDPC) code and Turbo code, often introduce a nonlinear mapping $v = f_1(x)$. The code word v is then modulated into a waveform $s = f_2(v) = f_2(f_1(x))$ which is then transmitted. Even when the channel noise is additive, namely the received waveform is $w = s + d$ where d is the channel noise, after the reverse process of demodulation and decoding, we have $y = g(w) = g(s + d) = g(f_2(f_1(x)) + d)$. As a result, the error term $g(f_2(f_1(x)) + d) - x$ in general is nonadditive and signal dependent. In addition, block and convolution coding schemes introduce temporally dependent noises. In our formulation, this aspect is reflected in dependent ϕ-mixing noises on ξ_n^{ij}. These will be detailed later.

For simplification on system derivations, we use first $d_n^{ij} = \xi_n^{ij}$ in this section. Let $\widetilde{\eta}_n$ and ξ_n be the l_s dimensional vectors that contain all $\widehat{x}_n^{ij}$ and ξ_n^{ij} in a selected order, respectively. Then, (2.2) can be written as $\widetilde{\eta}_n = H_1 x_n + \xi_n$, where H_1 is an $l_s \times r$ matrix whose rows are elementary vectors such that if the ℓth element of $\widetilde{\zeta}_n$ is $\widehat{x}^{ij}$ then the ℓth row in H_1 is the row vector of all zeros except for a "1" at the jth position. Each sensing link provides information $\delta_n^{ij} = x_n^i - \widehat{x}_n^{ij}$, an estimated difference between x_n^i and x_n^j. This information may be represented, in the same arrangement as $\widetilde{\eta}_n$, by a vector δ_n of size l_s containing all δ_n^{ij} in the same order as $\widetilde{\eta}_n$. δ_n can be written as $\delta_n = H_2 x_n - \widetilde{\eta}_n = H_2 x_n - H_1 x_n - \xi_n = H x_n - \xi_n$, where H_2 is an $l_s \times r$ matrix whose rows are elementary vectors such that if the ℓth element of $\widetilde{\zeta}(k)$ is $\widehat{x}^{ij}$ then the ℓth row in H_2 is the row vector of all zeros except for a "1" at the ith position, and $H = H_2 - H_1$. The reader is referred to [2] for basic matrix properties in graphs and to [39] for matrix iterative schemes. Due to network constraints, the information δ_n^{ij} can only be used by nodes i and j. When the control is linear, time invariant, and memoryless, we have $v_n^{ij} = \mu g_{ij} \delta_n^{ij}$ where g_{ij} is the link control gain on (i, j) and μ is a global scaling factor that will be used in state updating algorithms as the recursive stepsize. Let G be the $l_s \times l_s$ diagonal matrix that has g_{ij} as its diagonal element. In this case, the node control becomes $u_n = -\mu H' G \delta_n$. For convergence analysis, we note that μ is a global control variable and we may represent u_n equivalently as $u_n = -\mu(H'GHx_n - H'G\xi_n) = \mu(Mx_n + W\xi_n)$, with $M = -H'GH$ and $W = H'G$.

2.2 Convergence to Consensus

Under the link-based state control u_n^i, the state updating scheme (2.1) becomes

$$x_{n+1} = x_n - \mu H' G \delta_n. \tag{2.4}$$

Since $\mathbb{1}'M = 0$, $\mathbb{1}'W = 0$, $\mathbb{1}'x_{n+1} = \mathbb{1}'x_n = r\eta$ hold for all n, which is a natural constraint to the stochastic approximation algorithm. Starting at x_0, x_n is updated iteratively by using (2.4), which for the analysis is

$$x_{n+1} = x_n + \mu(Mx_n + W\xi_n). \tag{2.5}$$

Throughout the paper, the noise $\{\xi_n\}$ is allowed to be correlated, both spatially and temporally. We will assume the following conditions.

(A0) (1) All link gains are positive, $g_{ij} > 0$. (2) $\mathcal{G}$ contains a spanning tree.

(A1) The observation noise $\{\xi_n\}$ is a sequence of stationary ϕ-mixing sequence such that $E\xi_n = 0$, $E|\xi_n|^{2+\Delta} < \infty$ for some $\Delta > 0$, and that the mixing measure $\widetilde{\phi}_n$ satisfies $\sum_{k=0}^{\infty} \widetilde{\phi}_n^{\Delta/(1+\Delta)} < \infty$, where $\widetilde{\phi}_n = \sup_{A \in \mathcal{F}^{n+m}} E^{(1+\Delta)/(2+\Delta)} |P(A|\mathcal{F}_m) - P(A)|^{(2+\Delta)/(1+\Delta)}$, $\mathcal{F}^{\xi}_{<n} = \sigma\{\xi_n; k < n\}$, $\mathcal{F}^{\xi}_{\geq n} = \sigma\{\xi_n; k \geq n\}$.

Remark 2.1. Recall that a square matrix $\widetilde{Q} = (\widetilde{q}_{ij})$ is a generator of a continuous-time Markov chain if $\widetilde{q}_{ij} \geq 0$ for all $i \neq j$ and $\sum_j \widetilde{q}_{ij} = 0$ for each i. Also, a generator or the associated continuous-time Markov chain is irreducible if the system of equations $\begin{cases} \nu\widetilde{Q} = 0, \\ \nu\mathbb{1} = 1 \end{cases}$ has a unique solution, where $\nu = [\nu_1, \ldots, \nu_r] \in \mathbb{R}^{1\times r}$ with $\nu_i > 0$ for each $i = 1, \ldots, r$ is the associated stationary distribution.

Under (A1), the noise is generally unbounded but has bounded $(2+\Delta)$th moments. In addition, it is a sequence of correlated noise, much beyond the usual i.i.d. (independent and identically distributed) noise classes. A ϕ-mixing sequence has the property that the remote past and the distant future are asymptotically independent. The asymptotic independence is reflected by the condition on the underlying mixing measure.

Theorem 2.2. *Under Assumption* (A0), (1) M *has rank* $r-1$ *and is negative semi-definite.* (2) M *is a generator of a continuous-time Markov chain, and is irreducible.*

Proof. (1) Under the hypothesis, G is full rank, positive definite. Since $\mathcal{G}$ contains a spanning tree, by [2, Lemma 2.5.1], H has rank $n-1$. From the expression $M = -H'GH$, these imply that M is negative semi-definite and has rank $r-1$.

(2) By $M = -H'GH$, it can be readily verified that all off-diagonal elements of M are in the form of 0 or $g_{ij} > 0$. From $H\mathbb{1} = 0$, all row sums and column sums of M are zero. Consequently, M is a generator of a continuous-time Markov chain. Since M is of rank $r-1$ and $M\mathbb{1} = 0$, $\nu = \mathbb{1}'/r$ satisfies $\nu M = 0$ and $\nu\mathbb{1} = 1$, and is the unique nonnegative solution. Therefore, M is irreducible. □

Studying algorithm (2.5) is within the framework of standard stochastic approximation methods; see [20, Chapter 8]. Associated with the algorithm, there is a limit ordinary differential equation

$$\dot{x} = Mx. \tag{2.6}$$

Letting $Mx = 0$, we obtain the equilibria of (2.6). Since M is a generator of a continuous-time Markov chain, the equilibria of (2.6) constitute the set $\mathsf{Z} = \{z \in \mathbb{R}^r,\ z = c\mathbb{1}$ for any real number $c \in \mathbb{R}\}$. That is, the equilibria are the set of r-dimensional vectors spanned by the vector $\mathbb{1}$. When $c = 0$, we get the equilibrium point 0, so Z is the set of consensus. Convergence of the recursive algorithms is closely related to the associated ODE (2.6). To analyze algorithm (2.5), using the ODE methods [20], we take a continuous-time interpolation $x^\mu(t) = x_n$ for $t \in [\mu n, \mu n + \mu)$ and study the limit dynamics through the trajectories of differential equations whose stationary points belong to Z. Recall (see [20, p. 104]) that a set S is said to be locally stable in the sense of Liapunov if for each $\delta > 0$ there is a $\delta_1 > 0$ such that all trajectories starting in the δ_1-neighborhood $N_{\delta_1}(S)$ of S never leave the δ-neighborhood $N_\delta(S)$ of S. If the trajectories ultimately go to S, then S is said to be asymptotically stable in the sense of Liapunov. If this holds for all initial conditions, then the asymptotic stability is said to be global. Following from the standard line of argument of stochastic approximation [20] with the utilization of the structure of M matrix, we obtain the proposition below.

Proposition 2.3. *Consider the algorithm* (2.5) *together with the constraint*

$$\mathbb{1}' x_n = \eta r. \tag{2.7}$$

Under Assumptions (A0) *and* (A1)*, for any $t_\mu \to \infty$ as $\mu \to 0$, $x^\mu(\cdot + t_\mu)$ converges in probability to $\eta\mathbb{1}$.*

Sketch of Proof. We only highlight the main ideas. Consider first (2.6). Define $V : \mathbb{R}^r \mapsto \mathbb{R}$ by $V(x) = x'x/2$. Then $V(0) = 0$, $V(x) > 0$ for $x \neq 0$, and $V(x) \to$

∞ as $|x| \to \infty$. Moreover, the derivative of $V(x)$ along the solution of (2.6) is $(d/dt)V(x(t)) = x'(t)Mx(t)$. By Theorem 2.2, M is negative semi-definite which implies $(d/dt)V(x(t)) = x'(t)Mx(t) \leq 0$. The stationary points of the above ODE are given by the solutions to the equation $Mx = 0$. Since M is a generator, the stationary points to (2.6) is precisely the set Z. By the invariant set theorem (see for example, [20, p.104]), as $t \to \infty$, the solution to (2.6) converges to Z. That is, Z is a globally asymptotically stable set. Using the methods in [20, Chapter 8], we can show that $x^\mu(\cdot)$ converges weakly to $x(\cdot)$ such that $x(\cdot)$ is a solution of (2.6). Moreover, taking $t_\mu \to \infty$ as $\mu \to 0$, $x^\mu(\cdot + t_\mu)$ converges to the set Z in probability. Furthermore, since the intersection of Z and $\mathbb{1}'z = \eta r$ is the single point $x = \eta\mathbb{1}$, we obtain $x^\mu(\cdot + t_\mu)$ converges in probability to the unique consensus solution $\eta\mathbb{1}$. The desired result thus follows.

3 Asymptotic Optimality for Consensus-Type SA Algorithms using Iterate Averaging

The benefits of the iterate averaging algorithm can be summarized by the following items. (1) The difficulty of selecting a good stepsize sequence $\{\mu_n\}$ in application is a handicap, and iterate averaging alleviates this difficulty by providing a systematic approach. (2) With the use of a large stepsize, i.e. one going to zero slower than $O(1/n)$, the algorithm forces the estimates to move towards the true parameter more quickly. (3) Iterate averaging smoothes out the noise effect and reduces the "variance" of the noise. As a result, it gives the best convergence rate with the best scaling factor and the "smallest asymptotic covariance." Further insight on this can be found in [20, Chapter 11]. It can also be shown that this optimality is related to the well-known Cramér-Rao lower bound (see [25]).

Using such an idea in this chapter, we build algorithms using iterate averaging for the purpose of reaching consensus. Rather than dealing with well-known consensus algorithms, we treat general classes of noise that can cover many communication schemes as an integrated part of networked systems. Nevertheless, neither the rate of convergence nor the optimality of a consensus-type algorithm can be obtained directly from existing results in SA theory. The matrix $\widehat{H}$ in the above paragraph needs to be Hurwitz. However, for our consensus problem, the corresponding matrix M (to be precisely defined in the following section) is a generator of a continuous-time Markov chain, which has a zero eigenvalue that makes the existing results not applicable. To overcome this difficulty, we use the irreducibility of M, which indicates that apart from zero, all other eigenvalues have negative real parts. We use the ordinary differential equation (ODE) approach (see [20]) in our analysis. In lieu of working with the discrete iterates directly, we take a continuous-time interpolation of the iterates. Then using compactness, we can show that the resulting sequence of functions converges to a solution of the ODE.

3.1 Algorithms

Based on the discussion of last section, we propose a class of stochastic approximation algorithms. In consideration of extensive early work on consensus control, we shall go to the algorithms directly. For previous work on such algorithms, we refer the reader to the references in [47]. Suppose $x \in \mathbb{R}^r$ and $W \in \mathbb{R}^{r \times r_1}$, $\widehat{W} : \mathbb{R}^r \times \mathbb{R}^{r_1} \mapsto \mathbb{R}^r$. We begin by considering the following state updating algorithm

$$x_{n+1} = x_n + \mu_n M x_n + \mu_n W \xi_n + \mu_n \widehat{W}(x_n, \zeta_n), \tag{3.1}$$

together with the constraint

$$\mathbb{1}' x_n = \beta r, \tag{3.2}$$

where $\{\mu_n\}$ is a sequence of stepsizes, M is an irreducible generator of a continuous-time Markov chain (hence $\mathbb{1}'M = 0$ and rank $M = r - 1$), $\{\xi_n\}$ and $\{\zeta_n\}$ are noise sequences taking values in $\mathbb{R}^{r_1}$, β is the team average, and consensus control aims to control each team member's state towards β. For example, in computer load balancing problems, β is the average per-processor work load. Equal distribution of the total work load on multiple processors permits efficient utility of computing resources. In flight coordination of team UAVs, β may be the average speed of the team. In terms of the consensus control in this paper, the goal is to move the team in a uniform speed, without changing the team speed as a pack. The algorithm includes an additive noise as well as a non-additive noise. Therefore the solution is sufficiently general to include many practical senarios in the setup. The stepsize satisfies the following conditions: $\mu_n \geq 0$, $\mu_n \to 0$ as $n \to \infty$, and $\sum_n \mu_n = \infty$. Some commonly used stepsize sequences include $\mu_n = a/n$ and $\mu_n = a/n^\gamma$ for $0 < \gamma \leq 1$. Since the algorithm (3.1) is a stochastic approximation procedure, we can use the general framework in Kushner and Yin [20] to analyze the asymptotic properties. Before proceeding further, we make a remark. If we assume that $W\mathbb{1} = 0$ and $\widehat{W}(x,\zeta)\mathbb{1} = 0$ for each x and each ζ, then $\mathbb{1}'x_{n+1} = \mathbb{1}'x_n = r\beta$ hold for all n and for some $\beta \in \mathbb{R}$ (In the algorithms considered in the literature, one often begins with $\widehat{W} = 0$ and W having the condition mentioned above). Thus, in this case, the constraint $\mathbb{1}'x_n = r\beta$ is always satisfied by the algorithm structure.

(A1) The noise $\{\xi_n\}$ is a stationary, ϕ-mixing sequence such that $E\xi_n = 0$, $E|\xi_n|^{2+\Delta} < \infty$ for some $\Delta > 0$, and the mixing measure $\widetilde{\phi}_n$ satisfies $\sum_{k=0}^{\infty} \widetilde{\phi}_n^{\Delta/(1+\Delta)} < \infty$, where $\widetilde{\phi}_n = \sup_{A\in\mathcal{F}^{n+m}} E^{(1+\Delta)/(2+\Delta)}|P(A|\mathcal{F}_m) - P(A)|^{(2+\Delta)/(1+\Delta)}$, $\mathcal{F}_n = \sigma\{\xi_n; k < n\}$, $\mathcal{F}^n = \sigma\{\xi_n; k \geq n\}$.

(A2) (i) The noise sequence $\{\zeta_n\}$ is a stationary sequence that is uniformly bounded and ϕ-mixing with mixing measure $\widehat{\phi}_n$ such that for each $x \in \mathbb{R}^r$, $E\widehat{W}(x,\zeta_n) = 0$, and the mixing rate condition holds with $\widetilde{\phi}_n$ replaced by $\widehat{\phi}_n$. (ii) $\widehat{W}(\cdot, \zeta)$ is a continuous function for each ζ and $\left|\widehat{W}(x, \zeta)\right| \leq K(1 + |x|)$ for each $x \in \mathbb{R}^r$ and ζ. (iii) $\{\xi_n\}$, and $\{\zeta_n\}$ are mutually independent.

Recall that we have assumed that M is a generator of a continuous-time Markov chain and is irreducible. One of the consequences of this above assumption is that M has zero as an eigenvalue with multiplicity one and all other eigenvalues have negative real parts. Another distinct feature of M is that the null space of M is spanned by the vector $\mathbb{1} = (1, \ldots, 1)' \in \mathbb{R}^r$. This characteristic is precisely why we can reach consensus. As a consequence of (A2), ϕ-mixing implies that the noise sequences $\{\xi_n\}$ and $\widehat{W}(x, \zeta_n)$ for each fixed x are strongly ergodic [16, p. 488] implying that as $n \to \infty$, we have

$$\begin{aligned} &\frac{1}{n} \sum_{j=m}^{m+n-1} \xi_j \to 0 \ \ \text{w.p.1}, \\ &\frac{1}{n} \sum_{j=m}^{m+n-1} \widehat{W}(x, \zeta_j) \to 0 \ \ \text{w.p.1}. \end{aligned} \tag{3.3}$$

If we are only interested in weak convergence, then we only need $\frac{1}{n}\sum_{j=m}^{m+n-1} E_m \widehat{W}(x, \zeta_j) \to 0$ in probability, where E_m denotes the conditioning on the σ-algebra $\mathcal{F}_m = \{\xi_{j-1}, \zeta_{j-1} : j \leq m\}$.

Idea of Technical Development. To study the convergence of the algorithm using the stochastic approximation methods developed in [20] instead of working with the discrete-time iterations, we examine sequences defined in an appropriate function space. This will enable us to get a limit ordinary differential equation (ODE). The significance of the ODE is that the stationary points are exactly the true parameters we wish to estimate. We define

$$t_n = \sum_{j=0}^{n-1} \mu_j, \ m(t) = \max\{n : t_n \le t\}, \tag{3.4}$$

the piecewise constant interpolation $x^0(t) = x_n$ for $t \in [t_n, t_{n+1})$, and the shift sequence $x^n(t) = x^0(t + t_n)$. We shall outline the main steps involved below. We can first derive a preliminary estimate on the second moments.

Lemma 3.1. *Under* (A1) *and* (A2)*, for any* $0 < T < \infty$*,*

$$\sup_{n \le m(T)} E|x_n|^2 \le K \ \textit{and} \ \sup_{0 \le t \le T} E|x^n(t)|^2 \le K, \tag{3.5}$$

for some $K > 0$*, where* $m(\cdot)$ *is defined in (3.4).*

Proof. We only indicate the main ideas and leave most of the details out. Concerning the first estimate, because of the boundedness of the second moment $E|\xi_n|^2$, the condition $\sum_{j=1}^{\infty} \mu_j^2 < \infty$, the boundedness of the nonadditive noise $\widehat{W}(x, \zeta_n)$, and the linear growth of $\widehat{W}(\cdot, \zeta)$ for each ζ, we can derive

$$E|x_n| \le K + K \sum_{j=1}^{n} \mu_j E|x_j|^2. \tag{3.6}$$

Here and henceforth, K is used as a generic positive constant, whose values may change for different usage. After an application of Grownwall's inequality to (3.6), and then taking the supremum over all $n \le m(T)$, the first error bound is obtained. Likewise, we can obtain the second estimate. □

Theorem 3.2. *Under Assumptions (A1) and (A2), the iterates generated by the stochastic approximation algorithm (3.1) satisfy* $x_n \to \beta \mathbb{1}$ *w.p.1 as* $n \to \infty$.

Proof. We only present the main idea below. We show that $\{x^n(\cdot)\}$ is equicontinuous in the extended sense (see [20, p. 102] for a definition) w.p.1. To verify this, we note that by the argument in the first part of the proof in [43, Theorem 3.1],

$$\sum_{j=1}^{\infty} \mu_j \xi_j \text{ converges w.p.1 and}$$

$$\sum_{j=1}^{\infty} \mu_j \widehat{W}(x, \zeta_j) \text{ converges w.p.1 for a fixed } x.$$

Define $\Phi^0(t) = \sum_{j=1}^{m(t)-1} \mu_j [\xi_j + \widehat{W}(x, \zeta_j)]$ and $\Phi^n(t) = \Phi^0(t_n + t)$, where $m(\cdot)$ is defined in (3.4). Then we can show that for each $T > 0$ and $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\limsup_n \sup_{0 \le |t-s| \le \delta} |\Phi^n(t) - \Phi^n(s)| \le \varepsilon \quad \text{w.p.1.}$$

The above estimate together with the form of the recursion then implies that $x^n(\cdot)$ is equicontinuous in the extended sense. Next, we can extract a convergent subsequence, which will be denoted by $x^{n_\ell}(\cdot)$. Then the Arzela-Ascoli theorem concludes that $x^{n_\ell}(\cdot)$ converges to a function $x(\cdot)$ which is the unique solution (since the recursion is linear in x) of the ordinary differential equation (ODE)

$$\dot{x}(t) = Mx(t). \tag{3.7}$$

Owing to the law of large numbers, the noise is averaged out. What is the significance of the limit ODE? To answer this question, we set the right-hand side of

(3.7) equal to zero ($Mx = 0$). We then obtain the stationary point of the ODE. Since M is a generator of a continuous-time Markov chain and is irreducible, the solutions to $Mx = 0$ constitute precisely the null space of M. The null space of M is spanned by the vector $\mathbb{1}$. That is, the set can be represented by $\Gamma = \{\gamma : \gamma = \gamma_0 \mathbb{1}, \gamma_0 \in \mathbb{R}\}$. Moreover, from basic properties of Markov chains (see [50, Appendix A.1]), as $t \to \infty$, the solution $x(t)$ to (3.7) satisfies that $x(t)$ converges to the set Γ. That is, $\operatorname{dist}(x(t), \Gamma) \to 0$ as $t \to \infty$, where $\operatorname{dist}(\cdot,\cdot)$ is the usual distance function defined by $\operatorname{dist}(x, \Gamma) = \inf_{y\in\Gamma} |x - y|$. Consequently, as $n \to \infty$ and $q(n_\ell) \to \infty$, $x^{n_\ell}(\cdot + q(n_\ell)) \to \Gamma$.

Furthermore, the algorithm (3.1) together with $x_n'\mathbb{1} = r\beta$ leads to the desired conclusion. The equilibria of the limit ODE (3.7) and this constraint lead to the following system of equations

$$\begin{cases} Mx = 0 \\ \\ \mathbb{1}'x = r\beta. \end{cases} \tag{3.8}$$

The irreducibility of M then implies that (3.8) has a unique solution $x_* = \beta\mathbb{1}$. In fact, by defining an augmented matrix $M_a = \begin{pmatrix} M \\ \\ \mathbb{1}' \end{pmatrix} \in \mathbb{R}^{(r+1)\times r}$, the above system may be written as

$$M_a x = \begin{pmatrix} \\ 0 \\ \\ \\ r\beta \\ \\ \end{pmatrix} := b_a \in \mathbb{R}^{(r+1)\times 1}. \tag{3.9}$$

Note that $M_a' M_a$ has full rank owing to the irreducibility of M. Thus the solution of (3.9) can be written as $x_* = (M_a' M_a)^{-1} M_a' b_a = \beta \mathbb{1}$.

3.2 Asymptotic Efficiency

To improve the efficiency we average iterates, resulting in a two-stage stochastic approximation algorithm. The idea is that we first obtain a coarse approximation by using a relatively large stepsize, and then we refine the approximation by taking an iterate average. For definiteness and simplicity, we take $\mu_n = 1/n^\gamma$ for some $(1/2) < \gamma < 1$. The algorithm is given as follows:

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{n^\gamma} M x_n + \frac{1}{n^\gamma} W \xi_n + \frac{1}{n^\gamma} \widehat{W}(x_n, \zeta_n), \\ \overline{x}_{n+1} &= \overline{x}_n - \frac{1}{n+1} \overline{x}_n + \frac{1}{n+1} x_{n+1}. \end{aligned} \tag{3.10}$$

If we assume that $W\mathbb{1} = 0$ and $\widehat{W}(x, \zeta)\mathbb{1} = 0$ for each x and each ζ, then $\mathbb{1}' \overline{x}_n = r\beta$.

Theorem 3.3. *Suppose the conditions of Theorem* 4.4 *are satisfied. For iterates generated by algorithm (3.10)* (*together with the constraint $\mathbb{1}' x_n = r\beta$*)*, $x_n \to \beta \mathbb{1}$ w.p.1 as $n \to \infty$.*

Similar to what was alluded to in the introduction, the benefits of iterate averaging for the consensus algorithm include a faster approach to a neighborhood of the true parameter in its initial stage, a straightforward way of selecting the stepsize sequences, and the optimal convergence rate. To emphasize the dimension of the vector $\mathbb{1}$, we sometimes write $\mathbb{1}_\kappa$ for an integer κ in what follows. Since M has rank $r-1$, without loss of generality, assume that the first $r-1$ columns are independent. Partition the matrices M and W as

$$M = \begin{pmatrix} M_{11} & M_{12} \\ \\ M_{21} & M_{22} \end{pmatrix}, \quad W = \begin{pmatrix} W_{11} & W_{12} \\ \\ W_{21} & W_{22} \end{pmatrix} \tag{3.11}$$

where $M_{11} \in \mathbb{R}^{(r-1)\times(r-1)}$, $M_{12} \in \mathbb{R}^{(r-1)\times 1}$, $M_{21} \in \mathbb{R}^{(r-1)\times 1}$, $M_{22} \in \mathbb{R}^{1\times 1}$, and similarly for W_{ij}. Then M_{11} is nonsingular. Accordingly, we partition x_n, $\overline{x}_n$, and W as

$$x_n = \begin{pmatrix} \widetilde{x}_n \\ \\ x_{n,r} \end{pmatrix}, \quad \overline{x}_n = \begin{pmatrix} \Theta_n \\ \\ \overline{x}_{n,r} \end{pmatrix}, \tag{3.12}$$

respectively, with compatible dimensions as those of M. We will assume another condition. This condition essentially is a linearization of $\widehat{W}$ about the point x_*. Note that in (A3) below, $\widehat{W}_0 = \widehat{W}_x(x_*, \zeta)$. Partition $\widehat{W}_0$, ξ, and x_* similar to that of W and x, respectively. Our rate of convergence is a local analysis.

(A3) $\widehat{W}(x,\zeta) = \widehat{W}(x_*,\zeta) + \widehat{W}_0(x - x_*) + O(|x - x_*|^2)$.

Note that $x_{n,r} = \beta r - \mathbb{1}'_{r-1}\widetilde{x}_n$ and $\overline{x}_{n,r} = \beta r - \mathbb{1}'_{r-1}\Theta_n$. Using this together with the partition and (A3), we can convert the constrained stochastic approximation to an unconstrained one. That is, we can concentrate on the first $r-1$ components of x_n. It follows from (3.10) that

$$\begin{cases} \widetilde{x}_{n+1} = \widetilde{x}_n + \dfrac{1}{n^\gamma}\widetilde{M}\widetilde{x}_n + \dfrac{1}{n^\gamma}[\widehat{\xi}_n + \widehat{W}_1(x_*, \zeta_n)] \\ \qquad\qquad + \dfrac{1}{n^\gamma}[\widetilde{W}_0(\widetilde{x}_n - \widetilde{x}_*) + M_{12}\beta r + O(|\widetilde{x}_n - \widetilde{x}_*|^2)] \\ \Theta_{n+1} = \Theta_n - \dfrac{1}{n+1}\Theta_n + \dfrac{1}{n+1}\widetilde{x}_{n+1}, \end{cases} \tag{3.13}$$

where

$$\widetilde{M} = M_{11} - M_{12}\mathbb{1}'_{r-1},\ \widehat{\xi}_n = W_{11}\widetilde{\xi}_n + W_{12}\xi_{n,r},$$

$$\widetilde{W}_0 = \widehat{W}_{0,11} - \widehat{W}_{0,12}\mathbb{1}'_{r-1},$$

and $\widehat{W}_1(x_*, \zeta)$ is an $(r-1)$-vector consisting of the first $(r-1)$ components of $\widehat{W}(x_*, \zeta)$. Similar to Theorem 4.4, we can show that $\widetilde{x}_n \to \widetilde{x}_* = -\widetilde{M}^{-1}M_{12}\beta r$. Furthermore, we can show that $\Theta_n \to \widetilde{x}_*$ w.p.1 as $n \to \infty$. Note that when we define $\widetilde{z} = \widetilde{x} - \widetilde{x}_*$ and substitute it into (3.13), the term involving $M_{12}\beta r$ will disappear. To study the rates of convergence of x_n, we need only examine that of $\widetilde{x}_n$. To proceed, define

$$B_n(t) = \frac{1}{\sqrt{n}}\sum_{k=0}^{\lfloor nt \rfloor}[\widehat{\xi}_k + \widehat{W}(x_*, \zeta_n)], \quad t \in [0,1], \tag{3.14}$$

where $\lfloor t \rfloor$ denotes the integer part of t. We have the following lemma.

Lemma 3.4. *Under condition* (A2)*,* $B_n(\cdot)$ *converges weakly to* $B(\cdot)$ *an* $\mathbb{R}^{r-1}$*-dimensional Brownian motion such that* $EB(t) = 0$ *and covariance* $\Sigma_0 t$*, where*

$$\begin{aligned}\Sigma_0 = E[\widehat{\xi}_1\widehat{\xi}_1' + \widehat{W}_1(1)\widehat{W}_1'(1)] + \sum_{k=2}^{\infty} E[\widehat{\xi}_1\widehat{\xi}_k' + \widehat{\xi}_k\widehat{\xi}_1'] \\ + \sum_{k=2}^{\infty} E[\widehat{W}_1(k)\widehat{W}_1'(1) + \widehat{W}_1(1)\widehat{W}_1'(k)],\end{aligned} \tag{3.15}$$

where $\widehat{W}_1(k)$ *is an abbreviation of* $\widehat{W}_1(x_*, \zeta_k)$*.*

Proof. Note that $E\widehat{\xi}_n = 0$, and it is also a mixing sequence satisfying the conditions of (A2). The same observation holds for the sequence $\{\widehat{W}_1(x_*, \zeta_n)\}$. Next, $B_n(t) = B_n^1(T) + B_n^2(t)$, where $B_n^1(t)$ and $B_n^2(t)$ are rescaled sequences of sums of $\widehat{\xi}_k$'s and $\widehat{W}(k)$, respectively. It can be shown that (see [49, Lemma 3.1]), $B_n^i(\cdot)$ converges weakly to a Brownian motion $B^i(\cdot)$. Next, $\left\{\widehat{\xi}_n\right\}$ and $\left\{\widehat{W}_1(x_*, \zeta_n)\right\}$ are independent. The sum of $B^1(t) + B^2(t)$ is again a Brownian motion and with the desired covariance given by (3.15). □

Working with (3.13), we obtain

$$\begin{aligned}\widetilde{x}_{n+1} - \widetilde{x}_* = [\widetilde{x}_n - \widetilde{x}_*] + \frac{\Gamma}{n^\gamma}(\widetilde{x}_n - \widetilde{x}_*) \\ + \frac{1}{n^\gamma}[\widehat{\xi}_n + \widehat{W}_1(x_*, \zeta_n)] + \frac{1}{n^\gamma}(|\widetilde{x}_n - \widetilde{x}_*|^2),\end{aligned} \tag{3.16}$$

where $\Gamma = \widetilde{M} + \widetilde{W}_0$.

Define

$$A_{nj} = \begin{cases} \displaystyle\prod_{k=j+1}^{n} (I + \Gamma/k^{\gamma}), \; n \geq j+1; \\ \\ I; \; n = j. \end{cases}$$

Then for any $\kappa \geq 0$,

$$\widetilde{x}_{n+1} - \widetilde{x}_* = A_{n,\kappa-1}[\overline{x}_\kappa - \widetilde{x}_*]$$

$$+ \sum_{j=\kappa}^{n} \frac{1}{j^{\gamma}} A_{nj} O(|\widetilde{x}_n - \widetilde{x}_*|^2)$$

$$+ \sum_{j=\kappa}^{n} \frac{1}{j^{\gamma}} A_{nj} [\widehat{\xi}_j + \widehat{W}_1(j)],$$

and

$$\sqrt{n+1}[\Theta_{n+1} - \widetilde{x}_*]$$

$$= \frac{1}{\sqrt{n+1}} \sum_{k=1}^{\kappa-1} [\widetilde{x}_k - \widetilde{x}_*]$$

$$+ \frac{1}{\sqrt{n+1}} \sum_{k=\nu}^{n} \; A_{k,\kappa-1}[\widetilde{x}_\kappa - \widetilde{x}_*]$$

$$+ \frac{1}{\sqrt{n+1}} \sum_{k=\nu}^{n} \sum_{j=\kappa}^{k} \frac{1}{j^{\gamma}} A_{kj} O(|\widetilde{x}_j - \widetilde{x}_*|^2)$$

$$+ \frac{1}{\sqrt{n+1}} \sum_{k=\nu}^{n} \sum_{j=\kappa}^{k} \frac{1}{j^{\gamma}} A_{kj} [\widehat{\xi}_j + \widehat{W}_1(j)].$$

Note that

$$|A_{nj}| \leq \exp\left(-\lambda \sum_{k=j+1}^{n} k^{-\gamma} \right)$$

for some $\lambda > 0$. In what follows, we choose

$$\kappa = \kappa(n) = \left[\left(\frac{1-\gamma}{\lambda} \right) \ln \ln n \right]^{\frac{1}{1-\gamma}}.$$

To proceed, we define

$$\overline{B}_n(t) = \frac{\lfloor nt \rfloor}{\sqrt{n}} (\Theta_{\lfloor nt \rfloor + 1} - \widetilde{x}_*). \tag{3.17}$$

We next show that asymptotically, the "effective" term of the normalized error above is given by $-\Gamma^{-1} B_n(t)$.

Lemma 3.5. *In addition to the assumptions of* (A1)–(A3)*, assume* Γ *is a stable matrix* (*all of its eigenvalues have negative real parts*)*. Then for* $t \in [0, 1]$,

$$\overline{B}_n(t) = -\Gamma^{-1} B_n(t) + o(1), \quad \textit{where} \;\; o(1) \to 0$$

in probability uniformly in t *as* $n \to \infty$.

Remark 3.6. In the absence of the nonadditive noise, Γ becomes $\widetilde{M}$. The stability of $\widetilde{M}$ is verified by using the irreducibility of the generator M.

We are now ready to present the following theorem.

Theorem 3.7. *Under the conditions of Lemma* 3.5*, we have the following assertions:*

- $\overline{B}_n(\cdot)$ *converges weakly to* $\overline{B}(\cdot)$*, a Brownian motion with covariance* $\Gamma^{-1}\Sigma_0(\Gamma^{-1})'t$*;*
- $\widetilde{x}_n - \widetilde{x}_*$ *converges in distribution to a normal random variable with mean 0 and asymptotic covariance* $\Gamma^{-1}\Sigma_0(\Gamma^{-1})'t$.

Proof. We will be very brief. To prove the first part of the theorem, we need only to evaluate its covariance, which follows by the well-known Slutsky theorem. To obtain the second part, set $t = 1$ in part one. Using Lemma 3.5 and part of the theorem, the desired result follows. □

3.2.1 Matrix Stepsize and Optimality

The rate of convergence of algorithm (3.1) is equivalent to that of the first recursion in (3.13). This algorithm satisfies the sensing topology constraint and is strongly convergent, but the convergence speed of $\widetilde{x}_n$ is usually not optimal. Then, what is the optimal convergence speed? How can the optimal convergence speed be achieved? To compute the optimal convergence rates, we consider matrix step sizes, rather than the scalar μ_n. Recall that the rates of convergence of stochastic approximation algorithms are determined jointly by the scaling factor in the centered and scaled estimation errors, and its asymptotic covariance. Among the step sizes of the order $O(n^{-\gamma})$, $\gamma = 1$ gives the best order of convergence. Then, we need to find the best covariance matrix. One may use a matrix step size sequence $\mu_n = \widetilde{H}/n$, where $\widetilde{H}$ is a matrix-valued parameter to be used as a variable to optimize the asymptotic covariance. It is known that by choosing the matrix $\widetilde{H}$ suitably, it is possible to achieve optimal convergence speed [20, Chapter 10]. To study the rate of convergence, let us begin with

$$\widetilde{x}_{n+1} - \widetilde{x}_* = \widetilde{x}_n - \widetilde{x}_* - \mu_n[\Gamma(\widetilde{x}_n - \widetilde{x}_*) + \widehat{\xi}_n + \widehat{W}_1(n)],$$

with $\mu_n = \widetilde{H}/n$. Recall that we used the notation $\widehat{W_1}(n) = (\widehat{W_i}(x_*, \zeta_n) : i \le r-1)$. We can take a continuous time interpolation of $v_n = n^{1/2}(\widetilde{x}_n - \widetilde{x}_*)$. Using the approach in [20, Chapter 10], we obtain the limit of the interpolated (and shifted) sequence of v_n denoted by $V^n(\cdot)$. The limit is a solution of the following stochastic differential equation

$$dV = \left(\widetilde{H}\Gamma + \frac{I}{2}\right) V\,dt + \widetilde{H}\Sigma_0^{1/2} d\widetilde{B}(t),$$

where $\widetilde{B}(\cdot)$ is a standard Brownian motion and Σ_0 is the error covariance as given in (3.15). The asymptotic covariance as a function of $\widetilde{H}$ is then given by

$$\check{\Sigma}(\widetilde{H}) = \int_0^\infty \exp\left(D + \frac{I}{2}t\right) D\Sigma_0 D' \exp\left(D' + \frac{I}{2}t\right) dt,$$

where $D = \widetilde{H}\Gamma$. This can be alternatively represented as a solution to a Liapunov equation (or algebraic Riccati equation). Thus, $n^{1/2}(\widetilde{x}_n - \widetilde{x}_*)$ is asymptotically normal with mean zero and asymptotic covariance given by $\check{\Sigma}(\widetilde{H})$. To find the "smallest" asymptotic covariance, we either minimize $\check{\Sigma}(\widetilde{H})$ as a function of $\widetilde{H}$ or minimize the trace of the covariance. The optimal asymptotic covariance is given by

$$\Sigma_* = \Gamma^{-1}\Sigma_0(\Gamma')^{-1}. \tag{3.18}$$

However, as far as implementation is concerned, the matrix step size approach is usually impractical. The iterate averaging provides a viable alternative; see Theorem 3.7.

3.2.2 Optimal Convergence Rates

We now illustrate the optimality of the algorithms from another angle. For convergence speed analysis, let $e_n = x_n - x_*$. Decompose $e_n = [\widetilde{e}_n', e_{n,r}]'$ where $\widetilde{e}_n = \widetilde{x}_n - \widetilde{x}_*$.

Remark 3.8. *For simplicity, assume there is no nonadditive noise, i.e.,* $\widehat{W}(x,\zeta) = 0$*. Suppose that* $\{\xi_n\}$ *is a sequence of i.i.d. random variables with mean zero and covariance* $E\xi_n\xi_n' = \Sigma_0$*. Then the consensus errors satisfy that* $\sqrt{n}(\widetilde{x}_n - \widetilde{x}_*)$ *converges in distribution to a normal random variable with zero mean and covariance given by* $\Gamma^{-1}\Sigma_0(\Gamma^{-1})'$.

Note that the above result does not require any distributional information on the noise $\{\xi_n\}$, other than the zero mean and finite second moments. We now state the optimality of the algorithm when the density of ξ_1 is a smooth function.

Theorem 3.9. *Suppose that the noise* $\{\xi_n\}$ *is a sequence of i.i.d. noise with a density* f *that is continuously differentiable. Then the recursive sequence* $\{\widetilde{x}_n\}$ *is asymptotically efficient in the sense of the Cramér-Rao lower bound on* $E\widetilde{e}_n'\widetilde{e}_n$ *being asymptotically attained,* $nE\widetilde{e}_n'\widetilde{e}_n \to tr\left(\Gamma^{-1}\widetilde{\Sigma}_0(\Gamma^{-1})'\right)$ *as* $n \to \infty$.

The convergence speed and optimality of the iterate x_n are directly related to those of $\widetilde{x}_n$. Under the conditions of Theorem 3.9, the sequence $\{x_n\}$ from the algorithm (3.10) is asymptotically efficient in the sense of the Cramér-Rao lower bound on $Ee_n'e_n$ being asymptotically attained.

3.3 Illustrative Examples

In this section, we use an example to illustrate the benefits of employing the post-iterate averaging technique. The main advantages include more consistent control accuracy and faster convergence speeds.

Example 3.10. Since our algorithm maintains the total average of the node states at every step of control, $\sum_{i=1}^{r} x_1^i/r = \beta$ is a constant. The consensus error at the index n will be plotted by using the error norm $[(x_n - \beta\mathbb{1})'(x_n - \beta\mathbb{1})]^{1/2}$ the error norm.

In this example, we consider a networked system with five nodes. The initial states are $x_0^1 = 12$, $x_0^2 = 34$, $x_0^3 = 56$, $x_0^4 = 8$, $x_0^5 = 76$. The state average is $\beta = 37.2$, which will not change in the state update. Initial consensus error is $[(x_0 - \beta\mathbb{1})'(x_0 - \beta\mathbb{1})]^{1/2} = 57.94$.

The network interconnection is defined by the topology matrices

$$H_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The link control gain matrix is

$$G = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Consequently, from $H = H_2 - H_1$, we have

$$M = -H'GH = \begin{bmatrix} -6.4 & 2 & 0 & 2 & 2.4 \\ 2 & -2.6 & 0.6 & 0 & 0 \\ 0 & 0.6 & -3 & 2.4 & 0 \\ 2 & 0 & 2.4 & -6.4 & 2 \\ 2.4 & 0 & 0 & 2 & -4.4 \end{bmatrix}$$

and noise impact matrix $W = H' * G$ (with $a = 0.9997$) is

$$W = \begin{bmatrix} a & -a & 0 & 0 & 0 & 0 \\ 0 & a & -a & 0 & 0 & 0 \\ 0 & 0 & a & -a & 0 & 0 \\ 0 & 0 & 0 & a & -a & 0 \\ 0 & 0 & 0 & 0 & a & -a \end{bmatrix}.$$

The observations are corrupted by noises on each link, represented by the (vector) sequence $\{\xi_n\}$, whose elements are i.i.d. random variables with zero mean and variance $\sigma^2 = 40$. The noises are spatially independent, specifically observation noises on different links are independent. The SA algorithm is implemented with a fixed step size $\mu_n = 0.005$. The simulation runs for 400 steps.

Two algorithms are executed. The first one is the SA without post-iterate averaging. State trajectories of this algorithm are shown in the plots of Figure 1. The second algorithm adds the post-iterate averaging. The resulting state trajectories are illustrated in the left two plot of Figure 1. In both cases, the states converge to the team average. However, the SA with post-iterate averaging demonstrates improved convergence features with less volatility and faster convergence. This is consistent with our previous theoretical analysis. A further comparison of these two algorithms is shown in Figure 2 by their respective consensus error trajectories. The SA with post-iterate averaging displays faster convergence with less fluctuations.

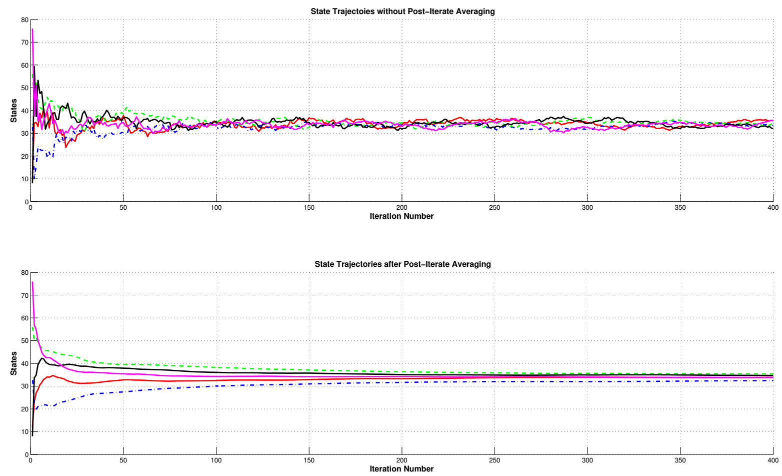


Figure 1: State trajectories of the two SA algorithms. Top plot: the standard SA algorithm. Bottom plot: The SA with added post-iterate averaging

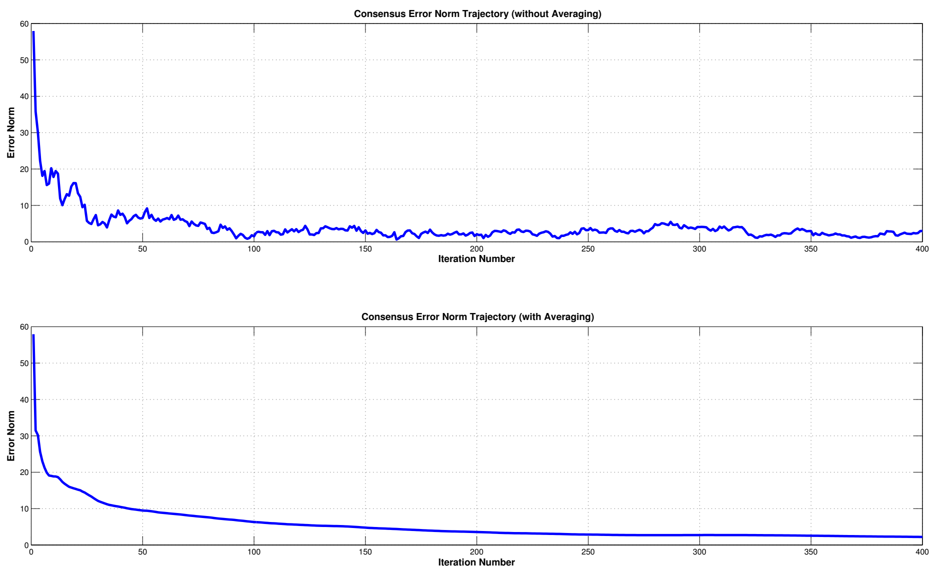


Figure 2: Comparison of consensus errors of the standard SA and the SA with post-iterate averaging

4 Time-Varying Network Topologies and Regime-Switching System

In this chapter, we carry out an extensive study of dealing with randomly regime-switching network topologies, whose parameters are time-varying and can be modeled by a discrete-time Markov Chain.

Switching network topologies were studied in [24, 26], and more recently in [12, 14]. This dissertation differs from the existing literature in several essential aspects. References [24, 26] do not use Markov formulations. In [12], the authors considered stochastic consensus over lossy wireless networks, in which the proposed measurement model has a random link gain, an additive noise, and a Markov lossy signal reception; arbitrary switching was also considered there. Reference [14] employs randomly switching Laplacian matrices together with observation noises that may be state dependent and Markovian. The Laplacian matrices share a common average. Its main approach is based on convergence of products of stochastic matrices. Thus, system analysis and consensus are established from the averaged network. We treat a more general Markov model and treat a much larger class of noises. In this dissertation, the graph is modulated by a discrete-time Markov chain. In addition to the traditional additive structure of the noise, we allow the noise to be nonadditive, correlated and non-Markovian. The function involved in the nonadditive noise can be time varying and depend on both the analog states and Markov chain states; see the remark section at the end of this paper. In lieu of examining the product of random matrices, our analysis is based on stochastic analysis of random processes. Thus far reaching results are obtained that better delineate the system dynamics and evolution. We establish convergence and rates of convergence of the algorithm, and study the intrinsic properties of the random dynamic systems involved. Interacting with consensus control strategies, we show that the limit system depends on relative speeds of the control and topology switching frequencies, and it may still be a stochastic system whose convergence is much harder to derive. By treating different rates of variation of the control and time-varying Markov parameter, our results depart from typical consensus control conclusions, initiate a multi-scale modeling and analysis, and potentially better reflect the needs of adjusting consensus control strategies in light of topology switching. Furthermore, the expanded classes of noises can cover many communication schemes.

The rest of the chapter is organized as follows. Section 4.1 begins with the algorithms under time-varying topologies and regime-switching. Section 4.2 proceeds with asymptotic properties concentrated on the case $\varepsilon = \mathcal{O}(\mu)$. Cases of $\varepsilon \ll \mu$, and $\mu \ll \varepsilon$ are discussed in section 4.3 and section 4.4. Finally, numberical examples are provided in section 4.5.

4.1 Algorithms

Suppose the network topology depends on a discrete-time Markov chain. In our setup, the graph can take m_0 possible values. The Markov chain is used to model, for example, capacity of the network, random environment, and other random factors such as interrupts and rerouting of communication channels etc. Thus $\mathcal{G}(\alpha_n) = \sum_{l=1}^{m_0} \mathcal{G}(l) I_{\{\alpha_n = l\}}$. To illustrate, suppose that initially, the Markov chain is at $\alpha_0 = i$. Then the graph takes the value $\mathcal{G}(i)$. At a random instance τ_1, the first jump of the Markov chain takes place so that $\alpha_{\tau_1} = j \neq i$, Then the graph switches to $\mathcal{G}(j)$ and holds that value for a random duration until the next jump of the Markov chain takes place.

To include topology switching and the extended noise class (2.3), the updating of network states is extended from (2.5) into

$$x_{n+1} = x_n + \mu M(\alpha_n) x_n + \mu \widetilde{W}(x_n, \alpha_n, \widetilde{\xi}_n), \tag{4.1}$$

where $\mu > 0$ is the step size of consensus control. For each $i \in \mathcal{M}$, $M(i)$ is a generator of a continuous-time Markov chain. The noise term $\widetilde{W}(\cdot,\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \times \mathbb{R}^r \mapsto \mathbb{R}^r$ is allowed to have the following general structure: for each $x \in \mathbb{R}^r$ and $i \in \mathcal{M}$,

$$\widetilde{W}(x, i, \widetilde{\xi}) = W(i)\xi + \widehat{W}(x, i, \zeta). \tag{4.2}$$

That is, it includes additive noise as well as nonadditive noise, When $W(i) = W$ a constant and $\widetilde{W} \equiv 0$, (4.2) reduces to the traditional additive noise. The nonadditive portion is a general nonlinear function of the analog state x, the Markov chain state $i \in \mathcal{M}$, as well as the noise source ζ_n. To state more explicitly dependence on ξ_n and ζ_n, in lieu of using the notation $\widetilde{\xi}_n$, we rewrite the algorithm as

$$x_{n+1} = x_n + \mu M(\alpha_n) x_n + \mu W(\alpha_n) \xi_n + \mu \widehat{W}(x_n, \alpha_n, \zeta_n) \tag{4.3}$$

in what follows. To proceed, we first give the assumptions needed for the noise sequence and the Markov chain α_n.

(A2) Assume the following conditions.

(a) α_n is a discrete-time Markov chain with a finite state space $\mathcal{M} = \{1, \dots, m_0\}$ representing the random environment and other random factors. The transition probability matrix of α_n is given by

$$P^\varepsilon = I + \varepsilon Q, \tag{4.4}$$

where $\varepsilon > 0$ is a small parameter, I is an $m_0 \times m_0$ identity matrix, and $Q = [q_{ij}] \in \mathbb{R}^{m_0 \times m_0}$ is the generator of a continuous-time Markov chain, (i.e., Q satisfies $q_{ij} \geq 0$ for $i \neq j$, $\sum_{j=1}^{m_0} q_{ij} = 0$ for each $i = 1, \dots, m_0$).

(b) The noise sequence $\{\xi_n\}$ is given in (A1).

(c) The $\{\zeta_n\}$ is a stationary sequence that is uniformly bounded such that for each $x \in \mathbb{R}^r$ and each $i \in \mathcal{M}$, $E\widehat{W}(x, i, \zeta_n) = 0$, and for any positive integer m,

$$\frac{1}{n} \sum_{j=m}^{m+n-1} E_m \widehat{W}(x, i, \zeta_j) \to 0 \text{ in probability,} \tag{4.5}$$

where E_m denotes the conditioning on the σ-algebra $\mathcal{F}_m = \{x_0, \alpha_j, \xi_{j-1}, \zeta_{j-1} : j \leq m\}$.

(d) $\widehat{W}(\cdot, i, \zeta)$ is a continuous function for each $i \in \mathcal{M}$ and each ζ and $|\widehat{W}(x, i, \zeta)| \leq K(1 + |x|)$ for each $x \in \mathbb{R}^r$, $i \in \mathcal{M}$, and ζ.

(e) $\{\alpha_n\}$, $\{\xi_n\}$, and $\{\zeta_n\}$ are mutually independent.

Remark 4.1. Concerning the assumptions above, we would like to make the following remarks. In our setup, $\{\zeta_n\}$ is another sequence of random variables. Suppose that it is a stationary mixing process then it is strongly ergodic, so for each fixed $x \in \mathbb{R}^r$, each $i \in \mathcal{M}$, and for any positive integer $m > 0$, the mixing and hence ergodicity implies that $\frac{1}{n}\sum_{j=m}^{m+n-1} \widehat{W}(x, i, \zeta_j) \to 0$ w.p.1. However, (4.5) is sufficient for this paper. Condition (d) indicates that $\widehat{W}(x, i, \zeta)$ grows at most linearly in x.

Although (4.3) is a stochastic approximation type algorithm, when switching topologies are present, its convergence is much harder to analyze. In the traditional setup of stochastic approximation problems, the limit or averaged system is an ordinary differential equation (ODE). Very often these limits are autonomous. Even if they are sometimes time inhomogeneous ordinary differential equations, these equations are non-random. As can be seen later, in certain problems treated here, the limit is no longer an ODE, but a randomly varying ODE subject to switching, owing to the Markov switching process. In the literature of stochastic approximation, the rate of convergence study is normally associated with a limit stochastic differential equation. In our case, some of the limits are Markovian-switching stochastic differential equations (i.e., switching diffusions [53]).

There are three possibilities concerning the relative sizes of ε and μ: (i) $\mu = \mathcal{O}(\varepsilon)$, (ii) $\varepsilon \ll \mu$, and (iii) $\mu \ll \varepsilon$. We first treat case (i) in detail, and then cover the other two cases. This idea also appears in related treatments of LMS-type algorithms under regime-switching dynamic systems, see [44, 45, 46]. In treating the three different cases, careful analysis is needed to examine convergence, stability, and related consensus issues. The next two sections will analyze the three cases.

4.2 Asymptotic Properties: $\varepsilon = \mathcal{O}(\mu)$

This section will concentrate on the case $\varepsilon = \mathcal{O}(\mu)$. For notational simplicity, in what follows, we simply consider $\varepsilon = \mu$, although general discussions do not incur further technical difficulties.

4.2.1 Basic Properties

To proceed, we first present a moment estimate for the recursive algorithm (4.3). In what follows and throughout the paper, we use K to denote a generic positive constant with the convention $K + K = K$ and $KK = K$.

Lemma 4.2. *Under Assumptions* (A1) *and* (A2), *for any* $0 < T < \infty$, $\sup_{0 \le n \le T/\varepsilon} E|x_n|^2 < K \exp(T) < \infty$ *where* $K > 0$ *is a constant.*

We are now ready to proceed to the convergence study of the algorithm. We need an additional assumption concerning the irreducibility of the generator Q. This is used when we are dealing with large time behavior ($t \to \infty$), which is concerned with the case that $\mu \to 0$, $n \to \infty$, and $\mu n \to \infty$.

(A3) The generator Q is irreducible.

4.2.2 Convergence

This subsection is devoted to obtaining asymptotic properties of the recursive algorithm (4.3). The first result concerns the property of the algorithm as $\varepsilon \to 0$ through an appropriate continuous-time interpolation. We define $x^\varepsilon(t) = x_n$, $\alpha^\varepsilon(t) = \alpha_n$, for $t \in [\varepsilon n, \varepsilon n + \varepsilon)$. Then $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot)) \in D([0,T] : \mathbb{R}^r \times \mathcal{M})$, which is the space of functions that are right continuous and have left limits endowed with the Skorohod topology [20, Chapter 7]. Before proceeding further, we first state a lemma that gives the weak convergence of the discrete iterates.

Lemma 4.3. *Under condition* (A2)*, the following claims hold:*

(a) *Denote* $p^\varepsilon_n = [P(\alpha^\varepsilon_n = 1), \ldots, P(\alpha^\varepsilon_n = m_0)]$ *and the* n*-step transition probability by* $(P^\varepsilon)^n$ *with* P^ε *given in* (4.4)*. Then*

$$\begin{aligned} p^\varepsilon_n &= p(t) + \mathcal{O}(\varepsilon + e^{-k_0 t/\varepsilon}), \\ (P^\varepsilon)^{n-n_0} &= \Xi(\varepsilon n, \varepsilon n_0) + \mathcal{O}(\varepsilon + e^{-k_0(n-n_0)}), \end{aligned} \tag{4.6}$$

where $p(t) \in \mathbb{R}^{1\times m_0}$ *and* $\Xi(t,t_0) \in \mathbb{R}^{m_0 \times m_0}$ *are the continuous-time probability vector and transition matrix satisfying*

$$\begin{aligned} \frac{dp(t)}{dt} &= p(t)Q,\ p(0) = p_0, \\ \frac{d\Xi(t,t_0)}{dt} &= \Xi(t,t_0)Q, \quad \Xi(t_0,t_0) = I, \end{aligned} \tag{4.7}$$

with $t_0 = \varepsilon n_0$ *and* $t = \varepsilon n$.

(b) $\alpha^\varepsilon(\cdot)$ *converges weakly to* $\alpha(\cdot)$*, which is a continuous-time Markov chain generated by* Q.

The proof of assertion (a) is essentially in that of Theorem 3.5 and Theorem 4.3

of [**?**], whereas the proof of (b) can be found in [52]; see also [51]. Thus the proof is omitted. We next obtain the weak convergence result.

Theorem 4.4. *Assume* (A1) *and* (A2). *Then* $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ *is tight in* $D([0,T] : \mathbb{R}^r \times \mathcal{M})$. *Moreover, as* $\varepsilon \to 0$, $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ *converges weakly to* $(x(\cdot), \alpha(\cdot))$ *that is a solution of the martingale problem with operator* $\mathcal{L}_1$. *For any* $f(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}$ *satisfying for each* $\alpha \in \mathcal{M}$, $f(\cdot, \alpha) \in C_0^1$ *(space of continuously differentiable functions with compact support),* $\mathcal{L}_1$ *is defined as follows:*

$$\mathcal{L}_1 f(x,i) = (\nabla f(x,i))' M(i)x + Qf(x,\cdot)(i), \quad i \in \mathcal{M}, \tag{4.8}$$

where $Qf(x,\cdot)(i) = \sum_{j=1}^{m_0} q_{ij} f(x,j)$.

Remark 4.5. An equivalent way of stating the martingale problem is to consider its associated differential equation. In this case, different from the traditional stochastic approximation problems, the limit dynamic system is not a deterministic differential equation, but a system of differential equations with random switching given by

$$\frac{dx(t)}{dt} = M(\alpha(t))x(t). \tag{4.9}$$

Proof of Theorem 4.4. The proof is divided into three steps. First, we prove that the tightness of $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$. Once the tightness is verified, we proceed to obtain the convergence using martingale problem formulation in the following three steps.

Step (i) Tightness. We treat the tightness of $\{x^\varepsilon(\cdot)\}$ and $\{\alpha^\varepsilon(\cdot)\}$ separately. The tightness of $\{\alpha^\varepsilon(\cdot)\}$ can be proved as in that of [51, Theorem 4.3]. Next we prove the tightness of $x^\varepsilon(\cdot)$, which is stated as a lemma below. In order to keep better flow of presentation, in what follows, we postpone some longer proofs of the technical results to the appendix.

Lemma 4.6. *Under the conditions of Theorem* 4.4*,* $\{x^\varepsilon(\cdot)\}$ *is tight in* $D([0,T]:\mathbb{R}^r)$*, which is the space of* $\mathbb{R}^r$*-valued functions that are right continuous and have the left limits, endowed with the Skorohod topology.*

Step (ii) By Lemma 4.6, $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ is tight. As a result, it is sequentially compact. Thus we can extract convergent subsequences. Next, it is important to ensure the limit of the convergent subsequence is unique. Thus we demonstrate that the solution for the martingale problem with operator $\mathcal{L}_1$ has a unique solution (unique in the sense of in distribution).

Lemma 4.7. *Under the conditions of Theorem* 4.4*, the martingale problem with operator* $\mathcal{L}_1$ *has a unique solution for each initial condition.*

Step (iii) To complete the proof, we characterize the limit process. Thus by virtue of the Prohorov theorem [20, p.229], we can extract a weakly convergent subsequence. For notational simplicity, we still denote the subsequence by $\{(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))\}$ with limit denoted by $(x(\cdot), \alpha(\cdot))$. To continue on our proof of the convergence result, we next show that the limit of $(x^\varepsilon(t), \alpha^\varepsilon(t))$ is a solution of the martingale problem with operator $\mathcal{L}_1$.

To characterize the limit property, we need to work with a continuously differentiable function with compact support $f(\cdot, \alpha)$ for each $\alpha \in \mathcal{M}$. Choose m_ε so that

$m_\varepsilon \to \infty$ but $\delta\varepsilon = \varepsilon m_\varepsilon \to 0$. Using the recursion (4.3),

$$
\begin{aligned}
& f(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - f(x^\varepsilon(t),\alpha^\varepsilon(t)) \\
&= \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} [f(x_{lm_\varepsilon+m_\varepsilon},\alpha_{lm_\varepsilon+m_\varepsilon}) - f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon})] \\
&= \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \Big\{ (\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))' \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} [M(\alpha_k)x_k \\
&\quad + W(\alpha_k)\xi_k + \widehat{W}(x_k,\alpha_k,\zeta_k)] \\
&\quad + [f(x_{lm_\varepsilon+m_\varepsilon},\alpha_{lm_\varepsilon+m_\varepsilon}) - f(x_{lm_\varepsilon+m_\varepsilon},\alpha_{lm_\varepsilon})] \Big\}.
\end{aligned}
\tag{4.10}
$$

The above representation will also be used in the rate of convergence study. We proceed to establish the next lemma, whose proof is provided in the appendix.

Lemma 4.8. *Under Theorem* 4.4*,* $(x^\varepsilon(\cdot),\alpha^\varepsilon(\cdot))$ *converges weakly to* $(x(\cdot),\alpha(\cdot))$*, which is the solution of the martingale problem with operator* $\mathcal{L}_1$*.*

Finally, piecing together the results obtained, the proof of the theorem is completed. □

4.2.3 Invariance Theorem

Note that the limit dynamics are not given by an ordinary differential equation, but rather a system of differential equations with Markov switching (4.9). How should we study the long-time behavior. It turns out a suitable way is the use of invariant set of the switched system. Following the discussion in [53, Chapter 9], recall that a

Borel measurable set $U \subset \mathbb{R}^r \times \mathcal{M}$ is invariant with respect to the solutions of (4.9) or simply, U is invariant with respect to the process $(x(t), \alpha(t))$ if $P((x(t), \alpha(t)) \in U$, for all $t \geq 0) = 1$, for any initial $(x, i) \in U$. That is, a process starting from U will remain in U w.p.1. We also need the notion of stability of sets in probability. They are defined naturally as follows. A closed and bounded set $K_c \subset \mathbb{R}^r$ is said to be stable in probability if for any $\delta > 0$ and $\rho > 0$, there is a $\delta_1 > 0$ such that starting from (x, i), $P(\sup_{t\geq 0} d(x(t), K_c) < \rho) \geq 1 - \delta$, whenever $d(x, K_c) < \delta_1$; asymptotically stable in probability if it is stable in probability, and moreover $P(\lim_{t\to\infty} d(x(t), K_c) = 0) \to 1$, as $d(x, K_c) \to 0$. In the above, we have used the usual distance function $d(x, D) = \inf(|x - y| : y \in D)$. We proceed to obtain the following result, whose proof is in the appendix.

Theorem 4.9. *Assume that for each $\alpha \in \mathcal{M}$, $M(\alpha)$ is irreducible. Under the conditions of Theorem* 4.4*, the following assertions hold.*

(i) *The set $\mathsf{Z} = span\{\mathbb{1}\}$ is an invariant set.*

(ii) *The set Z is asymptotically stable in probability.*

With the above proposition, we can further obtain the following result as a corollary of Theorem 4.4.

Corollary 4.10. *Assume the conditions of Theorem* 4.9*. In the recursive algorithm, we also use the constraint* (2.7)*. Then for any $t_\varepsilon \to \infty$ as $\varepsilon \to 0$, $x^\varepsilon(\cdot + t_\varepsilon)$ converges to the consensus solution $\eta\mathbb{1}$ in probability. That is for any $\delta > 0$, $\lim_{\varepsilon\to 0} P(|x^\varepsilon(\cdot + t_\varepsilon) - \eta\mathbb{1}| \geq \delta) = 0$.*

4.2.4 Normalized Error Sequences

This subsection is devoted to analyzing the rates of variations of scaled sequence of the errors and can be regarded as "rates of convergence" results. It is particularly interesting to derive results on the rate of convergence of x_n towards the limit $x(t)$. This can be examined through $x^\varepsilon(\varepsilon n) - x(\varepsilon n)$ for $n \leq \mathcal{O}(1/\varepsilon)$. For convenience, we work with a particular form of the nonadditive noise $\widehat{W}(x_n, \alpha_n, \zeta_n)$. Extension to more general case is presented in a later section.

(A2') Condition (A2) holds with the following modifications. Either $\widehat{W}(x, \alpha, \zeta) = \operatorname{diag}(x)\Psi(\alpha, \zeta)$ or $\widehat{W}(x, \alpha, \zeta) = x\psi_1(\alpha, \zeta)$, where $\Psi(\alpha, \zeta) : \mathcal{M} \times \mathbb{R}^r \mapsto \mathbb{R}^r$ and $\psi_1(\alpha, \zeta) : \mathcal{M} \times \mathbb{R}^r \mapsto \mathbb{R}$ such that $\Psi(\cdot, \cdot)$ (resp. $\psi_1(\alpha, \zeta)$) is a bounded function, and that for each fixed $\alpha \in \mathcal{M}$ and each positive integer m, (4.5) is replaced by

$$
\begin{aligned}
&\frac{1}{n} \sum_{j=m}^{m+n-1} E_m \Psi(\alpha, \zeta_j) \to 0 \ \text{ in probability,} \\
&\sum_{j=n}^{\infty} |E_n \Psi(\alpha, \zeta_j)| < \infty, \quad \text{or}
\end{aligned}
\tag{4.11}
$$

$$
\begin{aligned}
&\frac{1}{n} \sum_{j=m}^{m+n-1} E_m \psi_1(\alpha, \zeta_j) \to 0 \ \text{ in probability,} \\
&\sum_{j=n}^{\infty} |E_n \psi_1(\alpha, \zeta_j)| < \infty,
\end{aligned}
\tag{4.12}
$$

where $\operatorname{diag}(x) = \operatorname{diag}(x', \ldots, x')$.

For simplicity and definiteness, we use $\widehat{W}(x, \alpha, \zeta) = \operatorname{diag}(x)\Psi(\alpha, \zeta)$ in what follows. The argument for the use of $\psi_1(x, \alpha, \zeta)$ is exactly the same. To facilitate the analysis, we define an auxiliary sequence $\{y_n\}$ by

$$y_{n+1} = y_n + \varepsilon M(\alpha_n) y_n + \varepsilon \text{diag}(y_n)\Psi(\alpha_n, \zeta_n),\ \ y_0 = x_0. \tag{4.13}$$

This is a sequence having randomness only due to α_n and the non-additive noise. Define $y^\varepsilon(t) = y_n$ for $t \in [\varepsilon n, \varepsilon n + \varepsilon)$. Then a similar analysis to the proof of Theorem 4.4 yields the following result.

Lemma 4.11. *Under* (A2'), $y^\varepsilon(\cdot)$ *converges weakly to* $y(\cdot)$ *such that* $y(\cdot)$ *is a solution of the switching ordinary differential equation*

$$\dot{y}(t) = M(\alpha(t)) y(t). \tag{4.14}$$

Remark 4.12. Clearly, (4.14) is identical to the limit in Theorem 4.4. Compared with (4.13), (4.14) can be thought as an "averaged" system with the average interpreted in an appropriate sense.

To proceed, define

$$z_n = \frac{x_n - y_n}{\sqrt{\mu}} = \frac{x_n - y_n}{\sqrt{\varepsilon}} \ \ \text{since} \ \ \mu = \varepsilon. \tag{4.15}$$

Then it is readily verified that

$$z_{n+1} = z_n + \varepsilon M(\alpha_n) z_n + \sqrt{\varepsilon} W(\alpha_n)\xi_n + \varepsilon \text{diag}(z_n)\Psi(\alpha_n, \zeta_n). \tag{4.16}$$

We are in a position to study the asymptotic properties of the tracking error through weak convergence of appropriately interpolated sequence of z_n. Before proceeding further, we first obtain a second moment bound.

Lemma 4.13. *Assume that* (A1) *and* (A2') *hold. For any* $T < \infty$ *and for some* N_ε, $E \sup_{N_\varepsilon \le n \le T/\varepsilon} E|z_n|^2 = \mathcal{O}(1)$.

Note that normally, the bound obtained above can only be obtained after a "transient" period, which is reflected by the use of N_ε. Define $z^\varepsilon(t) = z_n$ for $t \in [(n - N_\varepsilon)\varepsilon, (n - N_\varepsilon)\varepsilon + \varepsilon)$.

Lemma 4.14. $\{z^\varepsilon(\cdot)\}$ *is tight on* $D([0, T] : \mathbb{R}^r)$.

Next extract a convergent subsequence $\{z^\varepsilon(\cdot)\}$. Without loss of generality, still denote the subsequence by $\{z^\varepsilon(\cdot)\}$ with limit $z(\cdot)$. For any $t, s > 0$,

$$\begin{aligned} z^\varepsilon(t+s) - z^\varepsilon(t) &= \varepsilon \sum_{l\delta_\varepsilon = t}^{t+s} \sum_{j=lm_\varepsilon}^{lm_\varepsilon + m_\varepsilon - 1} M(\alpha_j) z_j \\ &+ \varepsilon \sum_{l\delta_\varepsilon = t}^{t+s} \sum_{j=lm_\varepsilon}^{lm_\varepsilon + m_\varepsilon - 1} \operatorname{diag}(z_j) \Psi(\alpha_j, \zeta_j) \\ &+ \sqrt{\varepsilon} \sum_{l\delta_\varepsilon = t}^{t+s} \sum_{j=lm_\varepsilon}^{lm_\varepsilon + m_\varepsilon - 1} W(\alpha_j) \xi_j. \end{aligned} \tag{4.17}$$

The way to derive the limit is similar to that of Theorem 4.4. Keeping in mind that the limit will be a system of stochastic differential equations in which the switching process will come into play, we can then proceed to show that the limit is a solution of a martingale problem with a unique solution (in distribution). Baring this in mind, we will directly work with the sequence.

To proceed with the characterization of the limit process, define $\widehat{B}^\varepsilon(t) = \sqrt{\varepsilon} \sum_{j=N_\varepsilon}^{N_\varepsilon + t/\varepsilon - 1} \xi_j$. Then the mixing condition implies that $\widehat{B}^\varepsilon(\cdot)$ converges weakly to $\widehat{B}(\cdot)$, a Brownian motion with covariance $t\Sigma$, where Σ is given by

$$\Sigma = E\xi_0\xi_0' + \sum_{j=1}^{\infty} E\xi_j\xi_0' + \sum_{j=1}^{\infty} E\xi_0\xi_j'. \tag{4.18}$$

A proof of this fact may be found in [8, pp. 351–353]. Note that for any $j \in [lm_\varepsilon, lm_\varepsilon + m_\varepsilon)$ and $\varepsilon lm_\varepsilon \to v$, α_j can be replaced by $\alpha^\varepsilon(v)$,

$$\begin{aligned}
&\sqrt{\varepsilon}\sum_{l\delta_\varepsilon=t}^{t+s}\sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} W(\alpha_j)\xi_j \\
&= \sum_{l\delta_\varepsilon=t}^{t+s} W(\alpha^\varepsilon(v))[\widehat{B}^\varepsilon((l+1)\delta_\varepsilon) - \widehat{B}^\varepsilon(l\delta_\varepsilon)] \\
&\to \int_t^{t+s} W(\alpha(v))d\widehat{B}(v).
\end{aligned}$$

To summarize what have been obtained, we have the following theorem.

Theorem 4.15. *Under conditions* (A1), (A2'), *and* (A3), $(z^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ *converges to* $(z(\cdot), \alpha(\cdot))$ *such that* $z(\cdot)$ *is a solution of the following Markov regime-switching stochastic differential equation*

$$dz = M(\alpha(t))zdt + W(\alpha(t))d\widehat{B}(t). \tag{4.19}$$

4.3 Slowly Varying Markov Chains

Suppose that $\varepsilon \ll \mu$, where ε is the parameter appeared in the transition probability matrix of the Markov chain and μ is the stepsize of the algorithm (4.3). Intuitively, because the Markov chain changes so slowly, the time-varying parameter process is essentially a constant. We reveal the asymptotic properties of the recursive algorithm.

To facilitate the discussion and to simplify the notation, we take $\varepsilon = \mu^2$ in what follows.

Note that Lemma 4.3 still holds. We next use these to analyze the algorithm (4.3). As in the previous case, we can prove $\sup_{0\le n\le \mathcal{O}(1/\varepsilon)} E|x_n|^2 < \infty$. Define the piecewise constant interpolation $x^\mu(t) = x_n$ for $t \in [\mu n, \mu n+\mu)$. Then as in the previous section, we have $\{x^\mu(\cdot)\}$ is tight in $D([0,T],\mathbb{R}^r)$. We proceed to characterize its limit.

Since $x^\mu(\cdot)$ is tight, we can extract a convergent subsequence. For notational simplicity, still index the subsequence by μ with limit denoted by $x(\cdot)$. Note that

$$\begin{aligned} x^\mu(t+s) - x^\mu(t) &= \mu \sum_{j=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} [M(j)x_k + W(j)\xi_k \\ &\quad + \widehat{W}(x_k, j, \zeta_k)] I_{\{\alpha_k=j\}} \\ &= \sum_{j=1}^{m_0} \sum_{l\delta_\mu = t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [M(j)x_k + W(j)\xi_k \\ &\quad + \widehat{W}(x_k, j, \zeta_k)] I_{\{\alpha_k=j\}}. \end{aligned} \tag{4.20}$$

To figure out the limit, let us first look at

$$\begin{aligned} &\sum_{j=1}^{m_0} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} M(j)x_k I_{\{\alpha_k=j\}} \\ &\quad = \sum_{j=1}^{m_0} M(j) x_{lm_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} I_{\{\alpha_k=j\}} + o(1), \end{aligned} \tag{4.21}$$

where $o(1) \to 0$ in probability as $\mu \to 0$ uniformly in t. Next,

$$
\begin{aligned}
&\sum_{j=1}^{m_0} M(j) x_{lm_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} E_{lm_\mu} I_{\{\alpha_k=j\}} \\
&= M(\iota) x_{lm_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} P(\alpha_k = j | \alpha_{lm_\mu} = \iota) \\
&+ \sum_{j=1}^{m_0} c_j \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} P(\alpha_k = j | \alpha_{lm_\mu} = \iota)[I_{\{\alpha_{lm_\mu}=\iota\}} - 1] \\
&+ \sum_{j=1}^{m_0} c_j \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{j_1 \neq \iota} P(\alpha_k = j | \alpha_{lm_\mu} = j_1) I_{\{\alpha_{lm_\mu}=j_1\}},
\end{aligned}
\tag{4.22}
$$

where $c_j = M(j) x_{lm_\mu}/m_\mu$. For the last term above, we have

$$
\begin{aligned}
&E\Big|\frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{j_1 \neq \iota} P(\alpha_k = j | \alpha_{lm_\mu} = j_1) I_{\{\alpha_{lm_\mu}=j_1\}}\Big| \\
&= \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{j_1 \neq \iota} P(\alpha_k = j | \alpha_{lm_\mu} = j_1) \\
&\qquad \times P\{\alpha_{lm_\mu} = j_1 | \alpha_0 = \iota\} P(\alpha_0 = \iota) \\
&= \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{j_1 \neq \iota} P(\alpha_k = j | \alpha_{lm_\mu} = j_1) \\
&\qquad \times [\Xi_{\iota,j_1}(0, \varepsilon l m_\mu) + \mathcal{O}(\varepsilon + \exp(-lm_\mu))] P(\alpha_0 = \iota),
\end{aligned}
\tag{4.23}
$$

where $\Xi_{\iota,j_1}(0, \varepsilon l m_\mu)$ denotes the entry of the transition matrix (see Lemma 4.3) at the ιth row and j_1th column. Note that $\varepsilon l m_\mu \to 0$ as $\varepsilon \to 0$ since $\varepsilon = \mu^2$. Since $\Xi(0, \varepsilon l m_\mu) \to I$ the identity matrix, and for an off diagonal entry for $\iota \neq j_1$,

$\Xi_{\iota,j_1}(0,\varepsilon l m_\mu) \to 0$. In addition, $\sum_{lm_\mu=t/\delta_\mu}^{(t+s)/\delta_\mu} \delta_\mu \exp(-lm_\mu) \to 0$ as $\mu \to 0$. We can also show

$$E\left|\sum_{lm_\mu=t/\delta_\mu}^{(t+s)/\delta_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} P(\alpha_k = j|\alpha_{lm_\mu}=\iota) \times [I_{\{\alpha_{lm_\mu}=\iota\}} - 1]\right|^2 \to 0 \text{ as } \mu \to 0.$$

Thus, to find the limit in (4.22), it suffices to examine the term

$$M(\iota)x_{lm_\mu}\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} P(\alpha_k = j|\alpha_{lm_\mu}=\iota).$$

Then the martingale averaging techniques in [20] lead to

$$\mu\sum_{j=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu-1} M(j)x_k I_{\{\alpha_k=j\}} \to \int_t^{t+s} M(\iota)x(u)du. \tag{4.24}$$

Likewise, using detailed estimates similar arguments as in the previous section to handle the additive noise and nonadditive noise terms, we obtain

$$\begin{aligned}
&\mu\sum_{j=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu-1} W(j)\xi_k I_{\{\alpha_k=j\}} \to 0 \\
&\mu\sum_{j=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu-1} \widehat{W}(x_k,j,\zeta_k) I_{\{\alpha_k=j\}} \to 0.
\end{aligned} \tag{4.25}$$

Finally, since $\alpha_0 = \sum_{\iota=1}^{m_0} \iota I_{\{\alpha_0=\iota\}}$, we obtain the desired result with $M(\iota)$ in (4.24) replaced by $\sum_{\iota}^{m_0} p_\iota M(\iota)$. We summarize the discussions above into the following result.

Theorem 4.16. *Assume the conditions of Theorem* 4.4 *with the modification that the stepsize in* (4.3) *satisfies* $\varepsilon = \mu^2$. *Then* $x^\mu(\cdot)$ *converges weakly to* $x(\cdot)$, *which is a solution of the ordinary differential equation*

$$\dot{x}(t) = \sum_{\iota=1}^{m_0} p_\iota M(\iota)x(t). \tag{4.26}$$

Note that for each $\iota \in \mathcal{M}$, $M(\iota)$ is a generator of a continuous-time Markov chain. Since p_ι represents the initial probability distribution, it is nonnegative. As a result, $M_s = \sum_{\iota=1}^{m_0} p_\iota M(\iota)$ is also a generator of a continuous-time Markov chain. We have used M_s to signify that the generator correspond to slowly varying Markov chains. In view of the result in Section 2.2, we obtain the following corollary.

Corollary 4.17. *Assume the conditions of Theorem* 4.16 *and M_s is irreducible. In the recursive algorithm, we also use the constraint* (2.7)*. Then for any $t_\mu \to \infty$ as $\mu \to 0$, $x^\mu(\cdot + t_\mu)$ converges to the consensus solution $\eta \mathbb{1}$ in probability. That is for any $\delta > 0$, $\lim_{\mu \to 0} P(|x^\mu(\cdot + t_\varepsilon) - \eta \mathbb{1}| \geq \delta) = 0$.*

Remark 4.18. Note here we do not need the irreducibility of each of $M(\iota)$ but only the irreducibility of the average M_s. As was mentioned in the introduction, to avoid degeneracy, we require $\varepsilon > 0$. However, in fact, the result includes the degenerate case. If $\varepsilon = 0$, in lieu of a time-varying random parameter, there is only one "regime." Then there is only one M matrix. The requirement of M_s becomes that of M.

Furthermore, we may defined y_n as in (4.13) and define z_n as in (4.15). Then it can be shown that $\{z_n : n \geq N_\mu\}$ is tight. Define $z^\mu(t)$ to be the piecewise constant interpolation of z_n on $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$, then $z^\mu(\cdot)$ converges weakly to $z(\cdot)$ such that $z(\cdot)$ is the solution of the stochastic differential equation

$$dz(t) = \sum_{\iota=1}^{m_0} p_\iota M(\iota) z(t) dt + \sum_{\iota=1}^{m_0} p_\iota W(\iota) d\widehat{B}(t),$$

and $\widehat{B}(\cdot)$ is a Brownian motion with covariance Σt given in (4.18). We shall not dwell on the details here.

4.4 Fast Changing Markov Chains

This section takes up the issue that the Markov chain is fast varying comparing to the adaptation. By that, we mean $\mu \ll \varepsilon$. For concreteness of the discussion, we take a specific form of the stepsizes, namely, $\varepsilon = \mu^{1/2}$. Intuitively, the Markov chain vary relatively fast and can be thought of as a noise process. Eventually it is averaged out.

Consider again (4.3). Again, we can show that $x^\mu(\cdot)$ is tight. Then we can extract a convergent subsequence. For simplicity, still index the subsequence by μ with limit denoted by $x(\cdot)$. As in (4.20)–(4.22), choose a sequence m_μ such that $m_\mu \to \infty$ as $\mu \to 0$, but $\mu m_\mu \to 0$. Let us concentrate on the term

$$\begin{aligned}&\sum_{j=1}^{m_0} M(j) x_{lm_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} E_{lm_\mu} I_{\{\alpha_k=j\}} \\ &= \sum_{j=1}^{m_0} M(j) x_{lm_\mu} \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} P\{\alpha_k = j | \alpha_{lm_\mu}\}.\end{aligned} \tag{4.27}$$

For $\alpha_{lm_\mu} = i$,

$$P\{\alpha_k = j|\alpha_{lm_\mu}\} = \Xi_{ij}(\varepsilon l m_\mu, \varepsilon k) + \mathcal{O}(\varepsilon + \exp(-\kappa_0(k - lm_\mu))).$$

In view of (4.7) and noting $\varepsilon = \mu^{1/2}$ and the irreducibility of Q, we have $\Xi_{ij}(\varepsilon l m_\mu, \varepsilon k) = \nu_j + O\left(\exp\left(-\kappa_0 \frac{k-lm_\mu}{\sqrt{\mu}}\right)\right)$, where ν_j is the jth component of the stationary distribution $\nu = (\nu_1, \ldots, \nu_{m_0})$ associated with the generator Q of the corresponding continuous-time Markov chain. This indicates that $\Xi(s,t)$ can be approximated by a matrix $\mathbb{1}\nu$ with identical rows or what is equivalent, the initial state i is unimportant.

Thus detailed estimates yield that

$$
\begin{aligned}
&\mu \sum_{j=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} M(j) x_k I_{\{\alpha_k=j\}} \to \int_t^{t+s} \nu_j M(j) x(u) du, \\
&\mu \sum_{j=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} W(j) \xi_k I_{\{\alpha_k=j\}} \to 0 \\
&\mu \sum_{j=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu-1} \widehat{W}(x_k, j, \zeta_k) I_{\{\alpha_k=j\}} \to 0.
\end{aligned}
\tag{4.28}
$$

Thus we obtain the limit ordinary differential equation.

Theorem 4.19. *Assume the conditions of Theorem* 4.4 *with the modification that the stepsize in* (4.3) *satisfies* $\varepsilon = \mu^{1/2}$. *Then* $x^\mu(\cdot)$ *converges weakly to* $x(\cdot)$, *which is a solution of the ordinary differential equation*

$$
\dot{x}(t) = \sum_{j=1}^{m_0} \nu_j M(j) x(t). \tag{4.29}
$$

Similar to the slowly varying Markov chain case, for each $j \in \mathcal{M}$, $M(j)$ is a generator of a continuous-time Markov chain. The nonnegativity then yields that $M_f = \sum_{j=1}^{m_0} \nu_j M(j)$ is also a generator of a continuous-time Markov chain. We have used M_f to indicate that the generator correspond to fast varying Markov chains. The formulae (4.26) and (4.29) are similar in their appearance. The intuition behind is that for the slowly changing Markov chain case, the parameter is almost a constant resulting in a limit dynamic system "almost" like a constant parameter, whereas for the fast changing Markov chain, within a very short period of time, the system is replaced by an average with respect to the stationary distribution of the Markov chain. In view of the result in Section 2.2, we obtain the following corollary.

Corollary 4.20. *Assume that the conditions of Theorem* 4.19 *hold and that* M_f *is irreducible. In the recursive algorithm, we also use the constraint* (2.7)*. Then for any* $t_\mu \to \infty$ *as* $\mu \to 0$ *and for any* $\delta > 0$, $\lim_{\mu\to 0} P(|x^\mu(\cdot + t_\mu) - \eta \mathbb{1}| \geq \delta) = 0$.

Remark 4.21. The above result covers potentially interesting cases. We do not need the topology in each regime to be good (or irreducible), but only need the combined matrix M_f to have rank $r-1$. This will allow the possible loss of communications to happen that may create a topology that is not good on its own, but on average the combined network topologies provide sufficient linkage to achieve consensus.

Furthermore, we may defined y_n as in (4.13) and define z_n as in (4.15). Then it can be shown that $\{z_n : n \geq N_\mu\}$ is tight. Define $z^\mu(t)$ to be the piecewise constant interpolation of z_n on $t \in [(n-N_\mu)\mu, (n-N_\mu)\mu + \mu)$, then $z^\mu(\cdot)$ converges weakly to $z(\cdot)$ such that $z(\cdot)$ is the solution of the stochastic differential equation $dz(t) = \sum_{j=1}^{m_0} \nu_j M(j) z(t)dt + \sum_{j=1}^{m_0} \nu_j W(j) d\widehat{B}(t)$, and $\widehat{B}(\cdot)$ is a Brownian motion with covariance Σt given in (4.18) Again, the details are omitted here.

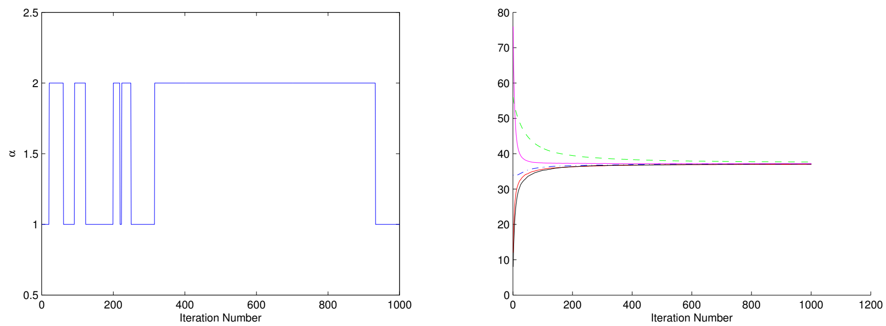
4.5 Illustrative Examples

This section presents several simulation examples. To obtain the desired consensus, we use $\sum_{i=1}^{r} x_0^i / r = \eta$. Then we call $(x_n - \eta \mathbb{1})'(x_n - \eta \mathbb{1})$ the consensus error variance at time n. We also term $(x_n - y_n)/\sqrt{\mu}$ the tracking error or scaled tracking error sequence.

Example 4.22. We consider the case that the Markov chain α_n has only 2 states, i.e., $\mathcal{M} = \{1, 2\}$. The probability transition matrix is $P^\varepsilon = I + \varepsilon Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \varepsilon \begin{pmatrix} -0.4 & 0.4 \\ 0.3 & -0.3 \end{pmatrix}$. For a given system, if the link gains $G_1 = \text{diag}(1, 0.3, 1.2, 4, 7, 10)$ and $G_2 = \text{diag}(2, 0.5, 1, 6, 9, 14)$ with regime-switching at two different states. Suppose the initial states are $x_0^1 = 12$, $x_0^2 = 34$, $x_0^3 = 56$, $x_0^4 = 8$, $x_0^5 = 76$. The state average is $\eta = 37.2$, which will not change in the state update. Initial consensus error is $(x_0 - \eta \mathbb{1})'(x_0 - \eta \mathbb{1}) = 3356.8$. Take $\varepsilon = 0.02$ and step size $\mu = \varepsilon = 0.02$. The updating algorithm runs for 1000 steps, and the stopped consensus error variance is $(x_{1000} - \eta \mathbb{1})'(x_{1000} - \eta \mathbb{1}) = 0.2355$. In what follows, we plot the Markov chain state trajectories, the system state trajectories. The consensus variance is shown to be fairly small.

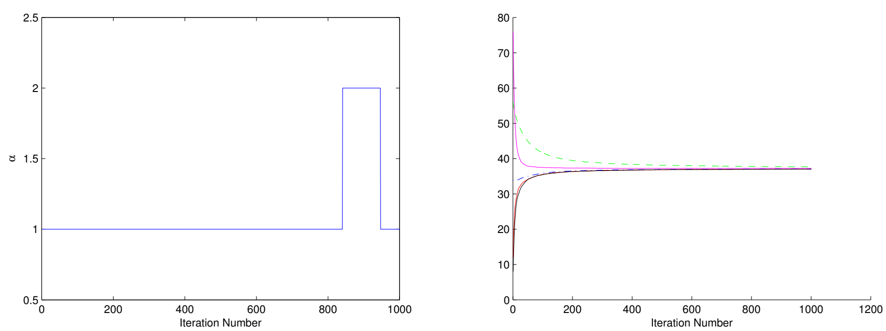
Example 4.23. Here we consider the case that the Markov chain changes very slowly compared with the adaptation stepsize. That is, $\varepsilon \ll \mu$. To be specific, suppose $\varepsilon = \mu^2$, where $\mu = 0.02$. The numerical results are showing in Figure 4. From the trajectory of the Markov chain, there is only one switching take place in the first 1000 iterations. The convergence of the consensus is also demonstrated.

Example 4.24. Here we consider the fast changing Markov chain $\mu \ll \varepsilon$. Specifically,



(a) Sample paths of Markov chain (b) System states trajectories: Values of x_n

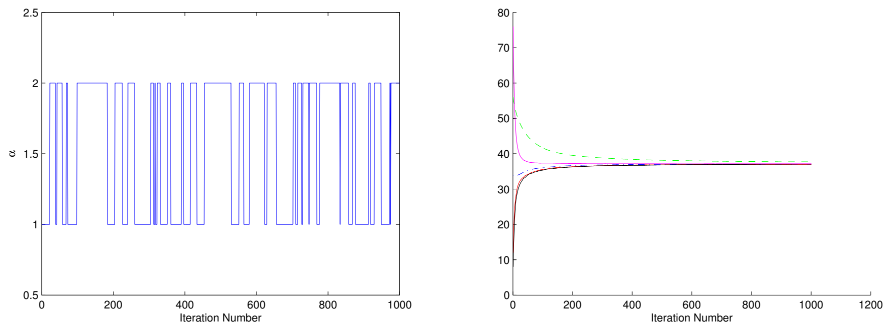
Figure 3: Trajectories of the case $\varepsilon = \mu = 0.02$: (Horizontal axes–discrete time or iteration numbers)



(a) Sample path of Markov chain (b) System states trajectories: Values of x_n

Figure 4: Slowly varying Markov parameter $\mu = 0.02$ and $\varepsilon = \mu^2$: (Horizontal axes–discrete time or iteration numbers)

we take $\mu = \varepsilon^2$ with $\mu = 0.02$ The corresponding trajectories plotted in Figure 5. The frequent Markov switching is clearly seen.



(a) Sample paths of Markov chain (b) System states trajectories: Values of x_n

Figure 5: Fast varying Markov chain $\varepsilon = \mu^{1/2}$ and $\mu = 0.02$: (Horizontal axes–discrete time or iteration numbers)

4.6 Proofs of Results

Proof of Lemma 4.2. Note that for any $0 < T < \infty$ and $0 \le n \le T/\varepsilon$, by the familiar Cauchy-Schwarz inequality,

$$\Big|\varepsilon \sum_{k=0}^{n} M(\alpha_k)x_k\Big|^2 \le \varepsilon^2 \Big(\sum_{k=0}^{n} 1^2\Big)\Big(\sum_{k=0}^{n} |M(\alpha_k)|^2|x_k|^2\Big), \tag{4.30}$$

so,

$$\varepsilon^2 E\left|\sum_{k=0}^{n} M(\alpha_k)x_k\right|^2 \le K\varepsilon \sum_{k=0}^{n} E|x_k|^2. \tag{4.31}$$

Also, $\varepsilon^2 E\left|\sum_{k=0}^{n} \widehat{W}(x_k, \alpha_k, \zeta_k)\right|^2 \le K\varepsilon \sum_{k=0}^{n} E|x_k|^2 + K$. Likewise,

$$\varepsilon^2 E\left|\sum_{k=0}^{n} W(\alpha_k)\xi_k\right|^2 \le K(\varepsilon n)^2 \le K. \tag{4.32}$$

Iterating on $E|x_n|^2$ with the use of (4.3), and using (4.31) and (4.32), we obtain

$$E|x_{n+1}|^2 \le (E|x_0|^2 + K) + K\varepsilon \sum_{k=0}^{n} E|x_k|^2.$$

An application of the Gronwall inequality then leads to $E|x_{n+1}|^2 \le K\exp(n\varepsilon) \le K\exp(\varepsilon(T/\varepsilon)) \le K\exp(T)$. Taking sup over n, the desired estimate follows. $\square$

Proof of Lemma 4.6. For any $\delta > 0$, let $t > 0$ and $s > 0$ such that $s \le \delta$, and $t, t+\delta \in [0,T]$. Note that

$$x^\varepsilon(t+s) - x^\varepsilon(t) = \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} M(\alpha_k)x_k$$

$$+\varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} W(\alpha_k)\xi_k + \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \widehat{W}(x_k,\alpha_k,\zeta_k).$$

In the above and hereafter, we use the convention that t/ε and $(t+s)/\varepsilon$ denote the corresponding integer parts, i.e., $\lfloor t/\varepsilon \rfloor$ and $\lfloor (t+s)/\varepsilon \rfloor$, respectively. However, for notational simplicity, in what follows, we will not use the floor function notation unless it is necessary.

Since α_k is a finite-state Markov chain, $|M(\alpha_k)|^2 \le \max_{i\in\mathcal{M}} |M(i)|^2 \le K$ and $|W(\alpha_k)|^2 \le \max_{i\in\mathcal{M}} |W(i)|^2 \le K$ a.s. Using the Cauchy-Schwarz inequality as in

(4.30) (with $\sum_{k=0}^{n}$ replaced by $\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}$) together with Lemma 4.2,

$$\begin{aligned}
&E|x^\varepsilon(t+s)-x^\varepsilon(t)|^2 \\
&\leq K\varepsilon^2 E\Big[\Big|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} M(\alpha_k)x_k\Big|^2+\Big|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} W(\alpha_k)\xi_k\Big|^2\Big] \\
&\quad +K\varepsilon^2 E\Big|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \widehat{W}(x_k,\alpha_k,\zeta_k)\Big|^2 \\
&\leq K\varepsilon s \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon} \sup_{t/\varepsilon\leq k\leq (t+s)/\varepsilon-1} E|x_k|^2+Ks^2\leq K\delta^2.
\end{aligned} \tag{4.33}$$

As a result, $\lim_{\delta\to 0}\limsup_{\varepsilon\to 0} E|x^\varepsilon(t+s)-x^\varepsilon(t)|^2=0$. The tightness of $\{x^\varepsilon(\cdot)\}$ follows from [19, p.47]. □

Proof of Lemma 4.7. Let $(x(t),\alpha(t))$ be a solution of the martingale problem with operator $\mathcal{L}_1$. We proceed to show that the solution is unique in the sense in distribution. Define

$$g(x,k)=\exp(\gamma' x+\gamma_0 k),\ \forall\gamma\in\mathbb{R}^r,\gamma_0\in\mathbb{R},k\in\mathcal{M}.$$

Consider $\psi_{jk}(t)=E[I_{\{\alpha(t)=j\}}g(x(t),k)],\ j,k\in\mathcal{M}$. It is readily seen that $\psi_{jk}(t)$ is the characteristic function associated with $(x(t),\alpha(t))$. By virtue of the Dynkin's formula

$$\psi_{j_0k_0}(t)-\psi_{j_0k_0}(0)-\int_0^t \mathcal{L}_1\psi_{j_0k_0}(s)ds=0. \tag{4.34}$$

Direct calculation also shows that

$$\mathcal{L}_1\psi_{j_0k_0}(s)=\gamma' M(k_0)x\psi_{j_0k_0}(s)+\sum_{j=1}^{m_0} q_{jj_0}\psi_{jk_0}(s). \tag{4.35}$$

Let $\psi(t) = (\psi_{\iota\ell}(t) : \iota \le m_0,\ \ell \le m_0)$. Combining (4.34) and (4.35), we then obtain

$$\psi(t) = \psi(0) + \int_0^t \psi(s) G ds, \tag{4.36}$$

where G is an $m_0 \times m_0$ matrix. Thus (4.36) is an ordinary differential equation with an initial condition $\psi(0)$. As a result, it has a unique solution. $\square$

Proof of Lemma 4.8. Our focus here is to characterize the limit below. By Skorohod representation [20, p. 230], with a slight abuse of notation, we may assume that $(x^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ converges to $(x(\cdot), \alpha(\cdot))$ w.p.1 and the convergence is uniform on any bounded time interval. To show that $(x(\cdot), \alpha(\cdot))$ is a solution of the martingale problem with operator $\mathcal{L}_1$, it suffices to show that for each $i \in \mathcal{M}$ and any $f(\cdot, i) \in C_0^1$, the class of functions that are continuously differentiable with compact support, $f(x(t), \alpha(t)) - f(x(0), \alpha(0)) - \int_0^t \mathcal{L}_1 f(x(s), \alpha(s)) ds$ is a martingale. To verify the martingale property, we need only show that for any bounded and continuous function $h(\cdot)$, any positive integer κ, any $t, s > 0$, and $t_i \le t$ with $i \le \kappa$,

$$\begin{aligned} &Eh(x(t_i), \alpha(t_i) : i \le \kappa)[f(x(t+s), \alpha(t+s)) \\ &\quad - f(x(t), \alpha(t)) - \int_t^{t+s} \mathcal{L}_1 f(x(u), \alpha(u)) du] = 0. \end{aligned} \tag{4.37}$$

To verify (4.37), we begin with the process indexed by ε. For notational simplicity, denote

$$\widetilde{h} = h(x(t_i), \alpha(t_i) : i \le \kappa),\ \widetilde{h}^\varepsilon = h(x^\varepsilon(t_i), \alpha^\varepsilon(t_i) : i \le \kappa). \tag{4.38}$$

The w.p.1 convergence (using the weak convergence and the Skorohod representation) together with the boundedness and the continuity of $f(\cdot)$ and $h(\cdot)$ yields that as $\varepsilon \to 0$,

$$E\widetilde{h}^{\varepsilon}[f(x^{\varepsilon}(t+s), \alpha^{\varepsilon}(t+s)) - f(x^{\varepsilon}(t), \alpha^{\varepsilon}(t))]$$

$$\to E\widetilde{h}[f(x(t+s), \alpha(t+s)) - f(x(t), \alpha(t))].$$

First, for the last term in (4.10), as $\varepsilon \to 0$,

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} [f(x_{lm_\varepsilon+m_\varepsilon}, \alpha_{lm_\varepsilon+m_\varepsilon}) - f(x_{lm_\varepsilon+m_\varepsilon}, \alpha_{lm_\varepsilon})] \\
&= \lim_{\varepsilon\to 0} \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} [f(x_{lm_\varepsilon}, \alpha_{lm_\varepsilon+m_\varepsilon}) - f(x_{lm_\varepsilon}, \alpha_{lm_\varepsilon})] \\
&= \int_t^{t+s} Qf(x(u), \cdot)(\alpha(u))du,
\end{aligned} \tag{4.39}$$

where $Qf(x,\cdot)(i)$ is as defined at the end of Theorem 4.4. Next, let us consider the term involving the noise. Since $h(x^{\varepsilon}(t_i), \alpha^{\varepsilon}(t_i) : i \le \kappa)$ is $\mathcal{F}_t^{\varepsilon}$ measurable, we can insert a conditional expectation with respect to $\mathcal{F}_t^{\varepsilon}$. Using assumption (A1) and [8, Corollary 7.2.4], for all $k \ge lm_\varepsilon$ and $t/\delta_\varepsilon \le l \le (t+s)/\delta_\varepsilon$,

$$\begin{aligned}
E|E_{lm_\varepsilon}\xi_k| &= E|E(\xi_k|\mathcal{F}_{lm_\varepsilon})| \\
&\le K\widetilde{\phi}_1^{\frac{1+\Delta}{2+\Delta}}(k - lm_\varepsilon)|\xi_k|_{2+\Delta} \\
&\le K\widetilde{\phi}^{\frac{1+\Delta}{2+\Delta}}(k - lm_\varepsilon)|\xi_1|_{2+\Delta},
\end{aligned}$$

where

$$\widetilde{\phi}_1(k) = \sup_{B\in\mathcal{F}^{n+m}} |P(B|\mathcal{F}_n) - P(B)|_1$$

and $|z|_q$ denotes the q-norm $E^{1/q}|z|^q$. Note that $W(\alpha_k) = \sum_{\ell=1}^{m_0} W(\ell) I_{\{\alpha_k=\ell\}}$. By the independence of $\{\alpha_k\}$ and $\{\xi_k\}$ and using (4.38),

$$
\begin{aligned}
\widetilde{e} &= |E\widetilde{h}^\varepsilon \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (\nabla f(x_{lm_\varepsilon}, \alpha_{lm_\varepsilon}))' W(\alpha_k)\xi_k| \\
&= |E\widetilde{h}^\varepsilon \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(x_{lm_\varepsilon}, \alpha_{lm_\varepsilon}))' \\
&\quad \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E_{lm_\varepsilon} W(\alpha_k) E_{lm_\varepsilon} \xi_k|.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\widetilde{e} &\leq \varepsilon E\widetilde{h}^\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{\iota,\ell=1}^{m_0} |\nabla f(x_{lm_\varepsilon}, \alpha_{lm_\varepsilon}))| \\
&\quad \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} P(\alpha_k = \ell | \alpha_{lm_\varepsilon} = \iota) |W(\ell)| |E_{lm_\varepsilon}\xi_k| I_{\{\alpha_{lm_\varepsilon}=\iota\}} \\
&\leq K\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E|E_{lm_\varepsilon}\xi_k| \\
&\leq K\varepsilon \left(\frac{t+s}{\delta_\varepsilon} - \frac{t}{\delta_\varepsilon}\right) \sum_{k=lm_\varepsilon}^{\infty} \widetilde{\phi}^{\frac{\Delta}{1+\Delta}}(k - lm_\varepsilon) \\
&\leq K\frac{\varepsilon}{\delta_\varepsilon} s \leq \frac{K}{m_\varepsilon} \to 0 \quad \text{as} \quad \varepsilon \to 0.
\end{aligned}
$$

For the nonadditive noise, by the continuity of $\widehat{W}(\cdot,\ell,\zeta)$, we have

$$\lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'\widehat{W}(x_k,\alpha,\zeta_k)$$

$$= \lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'$$

$$\times \sum_{\ell=1}^{m_0} \widehat{W}(x_{lm_\varepsilon},\ell,\zeta_k) I_{\{\alpha_k=\ell\}}.$$

Thus,

$$|E\widetilde{h}^{\varepsilon}\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'$$

$$\times \sum_{\ell=1}^{m_0} \widehat{W}(x_{lm_\varepsilon},\ell,\zeta_k) I_{\{\alpha_k=\ell\}}|$$

$$= |E\widetilde{h}^{\varepsilon}\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'$$

$$\times \sum_{\ell=1}^{m_0} E_{lm_\varepsilon}\widehat{W}(x_{lm_\varepsilon},\ell,\zeta_k) I_{\{\alpha_k=\ell\}}|.$$

As before, we can replace ε by $\delta_\varepsilon(1/m_\varepsilon)$. Using a partial summation,

$$
\begin{aligned}
&\sum_{\ell=1}^{m_0}\frac{1}{m_\varepsilon}\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E_{lm_\varepsilon}\widehat{W}(x_{lm_\varepsilon},\ell,\zeta_k)I_{\{\alpha_k=\ell\}}\\
&=\sum_{\ell=1}^{m_0}\sum_{\iota=1}^{m_0}\frac{1}{m_\varepsilon}\Big[P(\alpha_{lm_\varepsilon+m_\varepsilon-1}=\ell|\alpha_{lm_\varepsilon=\iota})\\
&\quad\times\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E_{lm_\varepsilon}\widehat{W}(x_{lm_\varepsilon},\ell,\zeta_k)\\
&\quad+\frac{1}{m_\varepsilon}\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}[P(\alpha_k=\ell|\alpha_{lm_\varepsilon=\iota})-P(\alpha_{k+1}=\ell|\alpha_{lm_\varepsilon=\iota})]\\
&\quad\times\sum_{j_1=lm_\varepsilon}^{k} E_{lm_\varepsilon}\widehat{W}(x_{lm_\varepsilon},\ell,\zeta_{j_1})\Big]I_{\{\alpha_{lm_\varepsilon}=\iota\}}.
\end{aligned}
\tag{4.40}
$$

Using (A2)(c) or (4.5), for all $lm_\varepsilon\le k\le lm_\varepsilon+m_\varepsilon-1$, $\frac{1}{m_\varepsilon}\sum_{j_1=lm_\varepsilon}^{k} E_{lm_\varepsilon}\widehat{W}(x_{lm_\varepsilon},\ell,\zeta_{j_1})\to 0$ in probability. Using the transition matrix $P^\varepsilon=I+\varepsilon Q$, for $lm_\varepsilon\le k\le lm_\varepsilon+m_\varepsilon-1$, $(I+\varepsilon Q)^{k-lm_\varepsilon}-(I+\varepsilon Q)^{k+1-lm_\varepsilon}=\mathcal{O}(\varepsilon)$. Using these estimates in (4.40), we obtain that

$$
E\widetilde{h}^\varepsilon\varepsilon\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon}\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}(\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'\widehat{W}(x_{lm_\varepsilon},\alpha,\zeta_k)\to 0.
$$

Next, we consider the term involving $M(\alpha_k)x_k$. We have

$$
\begin{aligned}
&\lim_{\varepsilon\to0}\varepsilon E\widetilde{h}^\varepsilon\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon}(\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}M(\alpha_k)x_k\\
&=\lim_{\varepsilon\to0}\varepsilon E\widetilde{h}^\varepsilon\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon}(\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}M(\alpha_k)x_{lm_\varepsilon}.
\end{aligned}
\tag{4.41}
$$

Thus, to get the desired limit, we need only examine the last line above. Let $\varepsilon lm_\varepsilon\to u$ as $\varepsilon\to0$. Then for all k satisfying $lm_\varepsilon\le k\le lm_\varepsilon+m_\varepsilon-1$, $\varepsilon k\to u$ since $\delta_\varepsilon\to0$. In addition, $\alpha_k = \alpha^\varepsilon(\varepsilon k)$. Thus

$$\begin{aligned}
&\lim_{\varepsilon\to 0} E\widetilde{h}^\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \frac{\delta_\varepsilon(\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'}{m_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} M(\alpha_k)x_{lm_\varepsilon} \\
&\quad = \lim_{\varepsilon\to 0} E\widetilde{h}^\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \frac{\delta_\varepsilon(\nabla f(x_{lm_\varepsilon},\alpha_{lm_\varepsilon}))'}{m_\varepsilon} \\
&\quad\quad \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} M(\alpha^\varepsilon(\varepsilon lm_\varepsilon))x_{lm_\varepsilon} \\
&\quad = E\widetilde{h}\int_t^{t+s} [\nabla f(x(u),\alpha(u))]'M(\alpha(u))x(u)du.
\end{aligned} \tag{4.42}$$

The desired result then follows. □

Proof of Theorem 4.9. To prove (i), we divide the time intervals by the associated switching times. Since we begin at $(x(0),\alpha(0)) = (x_0, i)$, we follow the dynamic system by considering its associated switching ordinary differential equation (4.9). Define τ_1 to be the first switching time, i.e., $\tau_1 = \inf\{t : \alpha(t) = i_1 \neq i\}$. Note that $x(t) = x(t,\omega)$, where $\omega \in \Omega$ is the sample point. Then in the interval $[0,\tau_1]$, for almost all ω, the system (4.9) is a system with constant matrix $M(i)$. Thus the solution can be represented by $x(t) = \exp(M(i)t)x_0$ for all $t \in [0,\tau_1]$. If $x_0 \in \mathsf{Z}$, i.e., $x_0 = c\mathbb{1}$ for some $c \in \mathbb{R}$, since $M(i)$ is a generator of a continuous-time Markov chain, and $\exp(M(i)t) = \sum_{k=0}^{\infty} \frac{(M(i)t)^k}{k!}$, $\exp(M(i)t)$ is orthogonal to $\mathbb{1}$. Thus $x(t) \in \mathsf{Z}$ for all $t \in [0,\tau_1]$. Now, define $\tau_2 = \inf\{t > \tau_1 : \alpha(t) = i_2 \neq i_1\}$. Similar as in the previous paragraph, we can show for all $t \in [\tau_1,\tau_2]$, $x(t) = \exp(M((i_1)(t-\tau_1))x(\tau_1)$ w.p.1. Moreover, $x(t) \in \mathsf{Z}$. Continue in this way. For any $T > 0$, consider $[0,T]$. Then

$0 < \tau_1 < \tau_2 \cdots < \tau_{n+1} \le T$, where τ_n is defined recursively such that $\alpha(\tau_n) = i_n$ and $\tau_{n+1} = \inf\{t > \tau_n : \alpha(t) = i_{n+1} \neq i_n\}$. Suppose that we have for all $t \le \tau_n$ $x(t) \in \mathsf{Z}$ w.p.1. Using the argument as before, we have $x(t) = \exp(M(i_n)(t-\tau_n))x(\tau_n)$ w.p.1 and $x(t) \in \mathsf{Z}$. Next, we work with the interval $[\tau_n, T]$, this establishes the first assertion.

To prove (ii), define $V(x) = x'x/2$. Note that since $V(x)$ is independent of the switching component, $\sum_{j=1}^{m_0} q_{ij} V(x) = 0$. Thus, $\mathcal{L}_1 V(x) = x'M(\alpha)x \le 0$, for each $\alpha \in \mathcal{M}$. The rest of the proof of the stability in probability of the set Z is similar in spirit to that of [53, Theorem 9.3]. We omit the details for brevity. □

Proof Lemma 4.13. The proof is carried out by using methods of perturbed Liapunov functions, which entitles to introduce small perturbations to a Liapunov function in order to make desired cancellation. We begin by introducing $V(z) = z'z/2$. Then

$$\begin{aligned} E_n V(z_{n+1}) - V(z_n) &= \varepsilon z_n' M(\alpha_n) z_n + \sqrt{\varepsilon} z_n' W(\alpha_n) E_n \xi_n \\ &+ \varepsilon z_n' \mathrm{diag}(z_n) \Psi(\alpha_n, \zeta_n) + \mathcal{O}(\varepsilon^2) V(z_n) \\ &+ \mathcal{O}(\varepsilon) |W(\alpha_n)|^2 E_n |\xi_n|^2 + \varepsilon^2 |\mathrm{diag}(z_n) \Psi(\alpha_n, \zeta_n)|^2, \end{aligned} \tag{4.43}$$

where E_n denotes the conditional expectation conditioned on the σ-algebra $\mathcal{F}_n = \{(x_j, \alpha_j) : j < n\}$.

Define

$$V_1^\varepsilon(z,n) = \sqrt{\varepsilon}\sum_{j=n}^{\infty} z' W(\alpha_n) E_n \xi_j,$$

$$V_2^\varepsilon(z,n) = \varepsilon\sum_{j=n}^{\infty} z' \mathrm{diag}(z) E_n \Psi(\alpha_n,\zeta_j).$$

Using (A1) and (A2'), we obtain

$$\begin{aligned} &|V_1^\varepsilon(z,n)| \le K|z|\sqrt{\varepsilon}\left|\sum_{j=n}^{\infty} E_n\xi_j\right| \le K\sqrt{\varepsilon}(V(z)+1) \\ &|V_2^\varepsilon(z,n)| \le K\varepsilon V(z). \end{aligned} \tag{4.44}$$

Moreover,

$$\begin{aligned} &E_n V_1^\varepsilon(z_{n+1},n+1) - V_1^\varepsilon(z_n,n) \\ &\quad = \mathcal{O}(\varepsilon)(V(z_n)+1) - \sqrt{\varepsilon} z_n' W(\alpha_n) E_n \xi_n \\ &E_n V_2^\varepsilon(z_{n+1},n+1) - V_1^\varepsilon(z_n,n) \\ &\quad = \mathcal{O}(\varepsilon^2)(V(z_n)+1) - \varepsilon z_n' \mathrm{diag}(z_n)\Psi(\alpha_n,\zeta_n). \end{aligned} \tag{4.45}$$

Define $V^\varepsilon(z,n) = V(z) + V_1^\varepsilon(z,n) + V_2^\varepsilon(z,n)$. Using (4.43) and (4.45), we obtain

$$\begin{aligned} &E_n V^\varepsilon(z_{n+1},n+1) - V^\varepsilon(z_n,n) \\ &\quad = \varepsilon z_n' M(\alpha_n) z_n + \mathcal{O}(\varepsilon) E_n|\xi_n|^2 + \mathcal{O}(\varepsilon)(V(z_n)+1). \end{aligned}$$

Thus for some $\kappa_1 > 0$,

$$E_n V^\varepsilon(z_{n+1}, n+1) \le (1-\kappa_1\varepsilon)V^\varepsilon(z_n, n) + \mathcal{O}(\varepsilon)\widetilde{\psi}_n, \tag{4.46}$$

where $\widetilde{\psi}_n$ satisfies $E|\widetilde{\psi}_n| < \infty$. Iterating on (4.46) and taking expectation, we obtain that

$$\begin{aligned} EV^\varepsilon(z_{n+1}, n+1) &\le (1-\kappa_1\varepsilon)^n EV^\varepsilon(z_0, 0) \\ &\quad + \mathcal{O}(\varepsilon)\sum_{j=0}^{n}(1-\kappa_1\varepsilon)^j. \end{aligned}$$

For sufficiently large n, $(1-\kappa_1\varepsilon)^n EV^\varepsilon(z_0, 0)$ can be made sufficiently small. Therefore, $\sup_{0\le n\le T/\varepsilon} EV^\varepsilon(z_n, n) = \mathcal{O}(1)$. Using the definition of $V^\varepsilon(z, n)$ together with (4.44), we also have $EV(z_n) = \mathcal{O}(1)$. The desired result thus follows. □

Proof of Lemma 4.14. As in the proof of Theorem 4.4, for any $\delta > 0$, $t, s > 0$ with $s < \delta$, (4.17) holds. Note that

$$\begin{aligned} &E\Big|\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} W(\alpha_j)\xi_j\Big|^2 \\ &= \varepsilon E \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \mathrm{tr}[E_{t/\varepsilon}W(\alpha_j)W'(\alpha_k)E_{t/\varepsilon}\xi_j\xi_k'] \\ &\le K\varepsilon\Big(\frac{t+s}{\varepsilon} - \frac{t}{\varepsilon}\Big) = Ks \le K\delta. \end{aligned}$$

The rest of the argument is similar to that of Lemma 4.6. A few details are omitted.

□

5 Concluding Remarks

This dissertation has been devoted to consensus type algorithms. In the first part, we developed a two-stage averaging algorithm and demonstrated its asymptotic optimality. In the setup of this part of the work, the topology is fixed. It would be a worthwhile effort to consider iterate averaging for algorithms with topology switching.

In the second part of the dissertation, we study asymptotic behavior of consensus-type algorithms for networked systems with randomly-switching topologies. Our results have demonstrated distinct convergence properties of three different scenarios. They are classified by the relative sizes of the Markov chain switching dynamics and the adaptation stepsizes. For convenience and notational simplicity, we have used the current setup. Several extensions and generalizations can be carried out. (a) In studying the rates of convergence, we used (A2'). This condition can be much relaxed. In lieu of (A2'), we can assume that $\widehat{W}(x,\alpha,\zeta) = \widehat{W}(x_c,\alpha,\zeta) + \nabla\widehat{W}(x_c,\alpha,\zeta)(x-x_c) + O(|x-x_c|^2)$, where $x_c = \eta\mathbb{1}$. In place of (4.11), we assume

$$\begin{aligned} &\frac{1}{n}\sum_{j=m}^{m+n-1} E_m \nabla\widehat{W}(x_c,\alpha,\zeta_j) \to 0 \ \text{ in probability,} \\ &\sum_{j=n}^{\infty} |E_n \nabla\widehat{W}(x_c,\alpha,\zeta_j)| < \infty. \end{aligned} \tag{5.1}$$

Proceeding in the same way as that of Theorem 4.15, we obtain the same limit switched stochastic differential equation. (b) So far, the noise sequences are correlated random processes. For convenience, we used mixing type of noise processes. All the development up to this point can be generalized to more complex x-dependent noise processes [20, Sections 6.6 and 8.4]. One possibility is to assume that the joint process $\{x_n, \alpha_{n-1}, \xi_{n-1}, \zeta_{n-1}\}$ is a Markov process and use the probabilistic structure of the joint process to carry out the analysis. (c) The nonadditive noise can be extended to incorporate time variations. That is, in lieu of a fixed function $\widehat{W}(x, \alpha, \zeta)$, we can treat time-varying $\widehat{W}_n(x, \alpha, \zeta)$. The main technique needed is a local average as in [20, p. 245, pp. 269-283]. For the extensions (a)–(c) mentioned above, the main line of developments will be along the line of the previous sections, but the notation will be more complex.

To conclude, this dissertation provides a class of general algorithms for consensus type of problems. This study opens new arenas for subsequent studies on consensus type control problems when time-varying and random dynamics of network systems are involved.

REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramniam, and E. Cayirci, A survey on sensor networks, *IEEE Commun. Mag.*, no. 8, pp. 102-114, Aug. 2002.

[2] R.A. Brualdi, H.J. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, 1991.

[3] David W. Casbeer and Randal W. Beard. Distributed information filtering using consensus filters. In *American Control Conference*, 2009.

[4] David W. Casbeer and Raymond W. Holsapple. Column generation for a uav assignment problem with precedence constraints. *International Journal of Robust and Nonlinear Control*, 21(12):1421–1433, 2011.

[5] J. Cortes and F. Bullo, Coordination and geometric optimization via distributed dynamical systems, *SIAM J. Control Optim.*, no. 5, pp. 1543-1574, May 2005.

[6] S. Haykin, *Digital Communications*, 4th ed., J. Wiley & Sons, 2001.

[7] K.L. Chung, On a stochastic approximation method, *Ann. Math. Statist.* **25** (1954), 463-483.

[8] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, J. Wiley, New York, NY, 1986.

[9] M. Gastpar and M. Vetterli. Source-channel communication in sensor networks, Springer Lecture Notes in Computer Science, vol. 2634, pp. 162-177, Apr. 2003.

[10] M. Huang and J.H. Manton, Stochastic approximation for consensus seeking: Mean square and almost sure convergence. Proc. 46th IEEE CDC Conference, New Orleans, LA, pp. 306-311, December 2007.

[11] 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.

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

[13] A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules, *IEEE Trans. Automat. Contr.*, vol. 48, pp. 988-1000, June 2003.

[14] S. Kar and J.M.F. Moura, Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise, *IEEE Trans. Signal Processing*, **57** (2009), no.1, 355–369.

[15] T.K. Moon, *Error Correction Coding, Mathematical Methods and Algorithms*, J. Wiley & Sons, 2005.

[16] S. Karlin and H.M. Taylor, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York, NY, 1975.

[17] Andrew Kwok and Raymond Holsapple. Approximate decentralized sensor fusion for bayesian search of a moving target. In *AIAA Infotech@Aerospace Conference*, 2011.

[18] H.J. Kushner and G. Yin, Asymptotic properties of distributed and communicating stochastic approximation algorithms, *SIAM J. Control Optim.*, **25** (1987), 1266–1290.

[19] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

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

[21] Y. Liu, K. Passino, and M.M. Polycarpou, Stability analysis of M-dimensional asynchronous swarms with a fixed communication topology, *IEEE Trans. Autom. Control*, vol. 48, no. 1, pp. 76-95, Jan. 2003.

[22] N. A. Lynch, *Distributed Algorithms*, Morgan Kaufmann Publishers, Inc., 1997.

[23] P. Travis Millet, David W. Casbeer, Travis Mercker, and Jacob L. Bishop. Multiagent decentralized search of a probability map with communication constraints. In *AIAA Guidance, Navigation and Control Conference*, 2010.

[24] L. Moreau, Stability of multiagent systems with time- dependent communication links, *IEEE Trans. Autom. Control*, vol. 50, no. 2, pp. 169-182, Feb. 2005.

[25] M.B. Nelson and R.Z. Khasminskii, *Stochastic Approximation and Recursive Estimation*, Amer. Math. Soc., Providence, RI, 1976.

[26] R. Olfati-Saber, J.A. Fax, and R.M. Murray. Consensus and cooperation in networked multi-agent systems, *IEEE Proc.*, vol. 95, no. 1, pp. 215-233, Jan. 2007.

[27] R. Olfati-Saber and R.M. Murray. Consensus problems in networks of agents with switching topology and time-delays, *IEEE Trans. Automatic Control*, vol. 49, pp. 1520-1533, Sep., 2004.

[28] P. Ogren, E. Fiorelli, and N.E. Leonard, Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment, *IEEE Trans. Autom. Control*, vol. 49, no. 8, pp. 1292-1302, Apr. 2005.

[29] F. Paganini, J. Doyle, and S Low. Scalable laws for stable network congestion control, *Proc. IEEE Conf. on Decision and Control, Orlando, FL*, Dec. 2001.

[30] B. T. Polyak, New method of stochastic approximation type, *Automation Remote Control* **7** (1991), 937–946.

[31] W. Ren and R.W. Beard. Consensus seeking in multiagent systems under dynamically changing interaction topologies, *IEEE Trans. Automat. Control*, vol. 50, no. 5, pp. 655-661, 2005.

[32] W. Ren, R. W. Beard, and D. B. Kingston. Multi-agent Kalman consensus with relative uncertainty. Proc. American Control Conf., Portland, OR, pp. 1865-1870, June 2005.

[33] C.W. Reynolds, Flocks, herds, and schools: a distributed behavioral model, *Computer Graphics*, **21**(4): 25-34, July 1987.

[34] D. Ruppert, Stochastic approximation, in *Handbook in Sequential Analysis*, B. K. Ghosh and P. K. Sen, eds., Marcel Dekker, New York, 1991, 503-529.

[35] S.S. Stanković and, M.S. Stankovic, and D.M. Stipanovic. Decentralized parameter estimation by consensus based stochastic approximation. *Automatic Control, IEEE Transactions on*, 56(3):531 –543, March 2011.

[36] J. Toner and Y. Tu, Flocks, herds, and schools: A quantitative theory of flocking, *Physical Review E*, **58** (4): 4828-4858, October 1998.

[37] T. Viseck, A. Czirook, E. Ben-Jacob, O. Cohen, and I. Shochet, Novel type of phase transition in a system of self-deriven particles, *Physical Review Letters*, **75** (6): 1226-1229, August, 1995.

[38] J.N. Tsitsiklis, D.P. Bertsekas, and M. Athans, Distributed asynchronous deterministic and stochastic gradient optimization algorithms, *IEEE Trans. Automat. Control*, **31**, no. 9, pp. 803-812, 1986.

[39] R.S. Varga, *Matrix Iterative Analysis*, Spring-Verlag, Berlin, 2000.

[40] L. Xiao, S. Boyd, and S. J. Kim, Distributed average consensus with least-mean-square deviation, *Journal of Parallel and Distributed Computing*, **67**, pp. 33-46, 2007.

[41] C.Z. Wei, Multivariate adaptive stochastic approximation, *Ann. Statist.* **15** (1987), 1115-1130.

[42] C. Xu and F. Lau, *Load Balancing in Parallel Computers: Theory and Practice*, Kluwer Academic Publishers, Boston, 1997.

[43] G. Yin, On extensions of Polyak's averaging approach to stochastic approximation, *Stochastics Stochastics Rep.*, **36** (1991), 245-264.

[44] G. Yin, C. Ion, and V. Krishnamurthy, How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump Markov sample paths, *Math. Programming*, Ser. B, **120** (2009), 67–99.

[45] G. Yin, V. Krishnamurthy, and C. Ion, Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization, *SIAM J. Optim.*, **14** (2004), 1187–1215.

[46] G. Yin and V. Krishnamurthy, Least mean square algorithms with Markov regime switching limit, *IEEE Trans. Automat. Control*, **50** (2005), 577–593.

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

[48] G. Yin, C.Z. Xu, and L.Y. Wang, Optimal remapping in dynamic bulk synchronous computations via a stochastic control approach, *IEEE Transactions on Parallel Distributed Systems*, **14** (2003), 51-62.

[49] G. Yin and K. Yin, Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms, *Stochastics Stochastics Rep.*, **47** (1994), 21–46.

[50] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.

[51] G. Yin and Q. Zhang, *Discrete-time Markov Chains: Two-time-scale Methods and Applications*, Springer, New York, NY, 2005.

[52] G. Yin, Q. Zhang, and G. Badowski, Discrete-time singularly perturbed Markov chains: Aggregation, occupation measures, and switching diffusion limit, *Adv. in Appl. Probab.*, 35(2), 449–476, 2003.

[53] G. Yin and C. Zhu, *Hybrid Switching Diffusions: Properties and Applications*, Springer, New York, 2010.

ABSTRACT

CONSENSUS-TYPE STOCHASTIC APPROXIMATION ALGORITHMS

by

YU SUN

August 2012

Advisor: Dr. G. George Yin

Major: Mathematics (Applied)

Degree: Doctor of Philosophy

This work is concerned with asymptotic properties of consensus-type algorithms for networked systems whose topologies switch randomly. The regime-switching process is modeled as a discrete-time Markov chain with a finite state space. The consensus control is achieved by designing stochastic approximation algorithms. In the setup, the regime-switching process (the Markov chain) contains a rate parameter $\varepsilon > 0$ in the transition probability matrix that characterizes how frequently the topology switches. On the other hand, the consensus control algorithm uses a stepsize μ that defines how fast the network states are updated. Depending on their relative values, three distinct scenarios emerge. Under suitable conditions, we show that when $0 < \varepsilon = \mathcal{O}(\mu)$, a continuous-time interpolation of the iterates converges weakly to a system of randomly switching ordinary differential equations modulated by a continuous-time Markov chain. In this case, a scaled sequence of tracking errors converges to a system of switching diffusion. When $0 < \varepsilon \ll \mu$, the network topology is almost non-switching during consensus control transient intervals, and hence the limit dynamic system is simply an autonomous differential equation. When $\mu \ll \varepsilon$, the Markov chain acts as a fast varying noise, and only its average is relevant, resulting in a limit differential equation that is an average with respect to the stationary

measure of the Markov chain. Simulation results are presented to demonstrate these findings.

By introducing a post-iteration averaging algorithm, this dissertation demonstrates that asymptotic optimality can be achieved in convergence rates of stochastic approximation algorithms for consensus control with structural constraints. The algorithm involves two stages. The first stage is a coarse approximation obtained using a sequence of large stepsizes. Then the second stage provides a refinement by averaging the iterates from the first stage. We show that the new algorithm is asymptotically efficient and gives the optimal convergence rates in the sense of the best scaling factor and "smallest" possible asymptotic variance. Numerical results are presented to illustrate the performance of the algorithm.

AUTOBIOGRAPHICAL STATEMENT

YU SUN

Education

- Ph.D. in Applied Mathematics, August 2012 (expected)
  Wayne State University, Detroit, Michigan
- M.A. in Mathematical Statistics, May 2011
  Wayne State University, Detroit, Michigan
- Graduate Certificate in College and University Teaching, August 2011
  Wayne State University, Detroit, Michigan
- M.S. in Applied Mathematics, March 2007
  Donghua University, Shanghai, China
- B.S. in Mathematical Education, July 2004
  Qingdao University, Shandong, China

Selected List of Awards

- The Robert and Nancy Irvan Endowed Mathematics Scholarship, Wayne State University, 2011
- The Katherine McDonald Award, Wayne State University, 2011
- The Excellence in Teaching Award of Graduate Employees Organizing Committee, Wayne State University, 2010
- Graduate Student Professional Travel Award, Department of Mathematics, Wayne State University, 2008

Selected List of Publications

1. Stochastic Recursive Algorithms for Networked Systems with Delay and Random Switching, SIAM Journal: Multiscale Modeling and Simulation, 9 (2011), 1087-1112. (with G. Yin and L. Y. Wang)
2. Asymptotic Properties of Consensus-type Algorithms for Networked Systems with Regime-switching Topologies, Automatica, 47 (2011), 1366-1378. (with G. Yin and L. Y. Wang)
3. An Alternative Model of Stochastic Volatility for Option Pricing: A Regime-switching Diffusion model under fast mean reversion, Dynamical Continuous and Discrete Impulse. System Ser. A Math. Anal., 16 (2009), Differential Equations and Dynamical Systems, suppl. S1, 29-35. (with G. Yin)
4. On the Linear Fractional Self-attracting Diffusion, J. Theoret. Probability. 21 (2008), no. 2, 502516; MR2391258 (2009d: 60112). (with L. Yan and Y. Lu)
5. Quantitative analysis of regional economic integration process affected by economy and trade cooperation between China and Japan, WSEAS Trans. Inf. Sci. Appl.,V6, No.1, 42-51 (2009). (with G. Wu and C. Gao)
6. Local Times of Some Processes Associated with Fractional Bessel Processes, Journal of Suzhou Institute of Technology (Nature) 24(2007), No. 1, 17-26. (with L. Yan)