1-1-2015

Stochastic Approximation Algorithms With Applications To Particle Swarm Optimization, Adaptive Optimization, And Consensus

Quan Yuan
Wayne State University,

STOCHASTIC APPROXIMATION ALGORITHMS WITH APPLICATIONS TO PARTICLE SWARM OPTIMIZATION, ADAPTIVE OPTIMIZATION, AND CONSENSUS

by

QUAN YUAN

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

2015

MAJOR: MATHEMATICS

Approved by:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

To my family and teachers

ACKNOWLEDGEMENTS

Over the past five years I have received support and encouragement from a great number of individuals. I would like to take this special opportunity to express my appreciation on all of them, especially to my advisor, Professor George Yin, for his endless support and care, spiritual guidance and academic training. Professor Yin's guidance has made this five years a thoughtful and rewarding journey. It is my great honor to be his student, and he is my lifetime role model.

I would like to thank my dissertation committee of Professor Boris Mordukhovich, Professor Kazuhiko Shinki, Professor Tze-Chien Sun, and Professor Wen Chen for their support over the past three years as I moved from rough ideas to a completed study. In addition, Professor Le Yi Wang also provided valuable support of the guidance and help for my PhD study.

During my graduate study, Professor Po Hu, Professor Alex Korostelev, Professor Guozhen Lu, Professors Bertram Schreiber, Professor Ualbai Umirbaev, and Professor Zhimin Zhang have taught me courses. I appreciate their help. Mrs. Patricia Bonesteel, Mrs. Tiana Bosley, Dr. John Breckenridge, Mrs. Mary Klamo, Mrs. Barbara Malicke, Dr. Choon-Jai Rhee, and Mrs. Joyce Wynn have trained me and supported me on teaching and many other aspects. I appreciate you all. I am also grateful to my friends Dan Ao, Yuehai Xu, Qi He, Guangliang Zhao, Xiaolong Han, Yayuan Xiao, Hongwei Mei, Xiaoyue Cui, Wei Ouyang, Yuan Tian, Ren Zhao, and Hailong Guo. Because of you all, my stay in Detroit has been colorful and memorable.

Finally, I thank my parents, Shuncai Yuan and Yuyan Shi, for the inspiring education and endless love they provided me. My thanks also go to my parents-in-law, Hongwei Yang

and Qinglan Yin, for taking care of my son when I am busy, and for supporting me in so many ways. My special thanks go to my wife Zhixin and my son Jacob, for sharing my hopes and dreams.

TABLE OF CONTENTS

CHAPTER 1 Introduction

1.1 Background and Main Issues

This dissertation focuses on stochastic approximation algorithms with some applications. In many real-world problems, the difficulties lie in the uncertainty in information. For example, in system identification the unknown system coefficients are estimated on the basis of input-output data of the control system; in adaptive control systems the adaptive control gain should be defined based on observation data in such a way that the gain asymptotically tends to the optimal one; researchers at a pharmaceutical form design laboratory experiments to extract the maximum information about the efficacy of a new drug, and more examples may be added to this list.

Many of these problems can be transformed to a root-seeking problem for an unknown function. To see this, let us consider a problem about estimating unknown parameters based on observation data containing information about the parameters. Let y_n denote the observation at time n i.e., the information available about the unknown parameters at time n. It can be assumed that the parameter under estimation denoted by x^* is a root of some unknown function $f(\cdot)$ with $f(x^*) = 0$. This is not a restriction, because, for example, $\|x - x^*\|^2$ may serve as such a function. Let x_n be the estimate for x^* at time n. Then the available information y_{n+1} at time $n + 1$ can formally be written as

$$y_{n+1} = f(x_n) \ + \varepsilon_{n+1},$$

where

$$\varepsilon_{n+1} = y_{n+1} - f(x_n).$$

Therefore, by considering y_{n+1} as an observation on $f(\cdot)$ at x_n with observation error ε_{n+1}, the problem has been reduced to seeking the root x^* of $f(\cdot)$ based on $\{y_n\}$.

If $f(\cdot)$ and its gradient can be observed without error at any desired values, then numerical methods such as Newton-Raphson method among others can be applied to solving the problem. However, this kind of methods cannot be used here, because in addition to the obvious problem concerning the existence and availability of the gradient, the observations are corrupted by errors which may contain not only the purely random component but also the structural error caused by inadequacy of the selected $f(\cdot)$.

Aiming at solving the stated problem, Robbins and Monro proposed the following recursive algorithm

$$x_{n+1} = x_n + a_n y_{n+1}, \qquad a_n > 0,$$

to approximate the sought-for root x^*, where a_n is the step size. This algorithm is now called the Robbins-Monro (RM) algorithm. Following this pioneer work of stochastic approximation, there have been a large amount of applications to practical problems and research works on theoretical issues.

At beginning, the probabilistic method was the main tool in convergence analysis for stochastic approximation algorithms, and rather restrictive conditions were imposed on both $f(\cdot)$ and $\{\varepsilon_n\}$. For example, it is required that the growth rate of $f(x)$ is not faster than linear as $\|x\|$ tends to infinity and $\{\varepsilon_n\}$ is a martingale difference sequence [1]. Though the linear growth rate condition is restrictive, as shown by simulation it can hardly be simply removed without violating convergence for RM algorithms. To weaken the noise conditions guaranteeing convergence of the algorithm, the ODE (ordinary differential equation) method was introduced in [3,79] and further developed in [78]. Since the conditions on noise required

by the ODE method may be satisfied by a large class of $\{\varepsilon_n\}$ including both random and structural errors, the ODE method has been widely applied for convergence analysis in different areas. In some applications people prefer to using constant step size

$$x_{n+1} = x_n + \eta y_{n+1},$$

where a constant $\eta > 0$ stands for a_n. The tool to deal with this situation, which was developed by Kushner [77, 90], called weak convergence method.

The development of stochastic approximation methods has been closely related to a wide range of applications in stochastic optimization, identification, adaptive control, estimation, detection, signal processing, management sciences, and many other related fields. As can be seen that many control and optimization tasks can be recast into a form that results in the use of stochastic approximation procedures. In this dissertation, we present three applications of stochastic approximation methods.

1.2 Outline of the Dissertation

The remainder of the dissertation is arranged as follows. In Chapter 2, we use stochastic approximation to analyze Particle Swarm Optimization (PSO) algorithm. We introduce four coefficients and rewrite the PSO procedure as a stochastic approximation type iterative algorithm. Then we analyze its convergence using weak convergence method. It is proved that a suitably scaled sequence of swarms converge to the solution of an ordinary differential equation. We also establish certain stability results. Moreover, convergence rates are ascertained by using weak convergence method. A centered and scaled sequence of the estimation errors is shown to have a diffusion limit.

In Chapter 3, we study a class of stochastic approximation algorithms with regime switching which is modulated by a discrete Markov chain having countable state spaces and two-time-scale structures. In the algorithm, the increments of a sequence of occupation measures are updated using constant step size. It is demonstrated that least squares estimates from the tracking errors can be developed. Under the assumption that the adaptation rates are of the same order of magnitude as that of times-different parameter, it is proven that the continuous-time interpolation from the iterates converges weakly to some system of ordinary differential equations (ODEs) with regime switching, and that a suitably scaled sequence of the tracking errors converges to a system of switching diffusion. This work is an extension of the work in [92].

In Chapter 4, we developed asynchronous stochastic approximation (SA) algorithms for networked systems with multi-agents and regime-switching topologies to achieve consensus control. There are several distinct features of the algorithms studied in the dissertation. (1) In contrast to the most existing consensus algorithms, the participating agents compute and communicate in an asynchronous fashion without using a global clock. (2) The agents compute and communicate at random times. (3) The regime-switching process is modeled as a discrete-time Markov chain with a finite state space. (4) The functions involved are allowed to vary with respect to time hence nonstationarity can be handled. (5) Multi-scale formulation enriches the applicability of the algorithms. In the setup, the switching process contains a rate parameter $\varepsilon > 0$ in the transition probability matrix that characterizes how frequently the topology switches. The algorithm uses a step-size μ that defines how fast the network states are updated. Depending on their relative values, three distinct scenarios emerge. Under suitable conditions, it is shown that 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, or to a system of differential equations (an average with respect to certain measure). In addition, a scaled sequence of tracking errors converges to a switching diffusion or a diffusion. Simulation results are presented to demonstrate these findings.

CHAPTER 2 Analyzing Convergence and Rates of Convergence of Particle Swarm Optimization Algorithms

2.1 Introduction

Recently, optimization using particle swarms have received considerable attention owing to the wide range of applications from networked systems, multi-agent systems, and autonomous systems. Particle swarming refers to a computational method that optimizes a problem by trying recursively to improve a candidate solution with respect to a certain performance measure. Swarm intelligence from bio-cooperation within groups of individuals can often provide efficient solutions for certain optimization problems. When birds are searching food, they exchange and share information. Each member benefits from all other members owing to their discovery and experience based on the information acquired locally. Then each participating member adjusts the next search direction in accordance with the individual's best position currently and the information communicated to this individual by its neighbors. When food sources scattered unpredictably, advantages of such collaboration was decisive. Inspired by this, Kennedy and Eberhart proposed a particle swarm optimization (PSO) algorithm in 1995 [26]. A PSO procedure is a stochastic optimization algorithm that mimics the foraging behavior of birds. The search space of the optimization problem is analogous to the flight space of birds. Using an abstract setup, each bird is modeled as a particle (a point in the space of interest). Finding the optimum is the counterpart of searching for food. A PSO can be carried out effectively by using an iterative scheme. The PSO algorithm simulates social behavior among individuals (particles) "flying" through a multidimensional search space, where each particle represents a point at the intersection of all search dimensions.

The particles evaluate their positions according to certain fitness functions at each iteration. The particles share memories of their "best" positions locally, and use the memories to adjust their own velocities and positions. Motivated by this scenario, a model is proposed to represent the traditional dynamics of particles.

To put this in a mathematical form, let $F : \mathbb{R}^D \to \mathbb{R}$ be the cost function to be minimized. If we let M denote the size of the swarm, the current position of particle i is denoted by X^i $(i = 1, 2, \ldots, M)$, and its current velocity is denoted by v^i. Then, the updating principle can be expressed as

$$\begin{aligned} v_{n+1}^{i,d} &= v_n^{i,d} + c_1 r_{1,n}^{i,d}[\mathrm{Pr}_n^{i,d} - X_n^{i,d}] + c_2 r_{2,n}^{i,d}[\mathrm{Pg}_n^{i,d} - X_n^{i,d}], \\ X_{n+1}^{i,d} &= X_n^{i,d} + v_{n+1}^{i,d}, \end{aligned} \tag{2.1}$$

where $d = 1, \ldots, D$; $r_1^{i,d} \sim U(0,1)$ and $r_2^{i,d} \sim U(0,1)$ represent two random variables uniformly distributed in $[0,1]$; c_1 and c_2 represent the acceleration coefficients; Pr_n^i represents the best position found by particle i up to "time" n, and Pg_n^i represents the "global" best position found by particle i's neighborhood Π_i, i.e.,

$$\begin{aligned} \mathrm{Pr}_n^i &= \arg \min_{1 \le k \le n} F(X_k^i), \\ \mathrm{Pg}_n^i &= \mathrm{Pr}_n^{j^*}, \text{ where } j^* = \arg\min_{j \in \Pi_i} F(\mathrm{Pr}_n^i). \end{aligned} \tag{2.2}$$

In artificial life and social psychology, v_n^i in (2.1) is the velocity of particle i at time n, which provides the momentum for particles to pass through the search space. The $c_1 r_{1,n}^{i,d}[\mathrm{Pr}_n^{i,d} - X_n^{i,d}]$ is known as the "cognitive" component, which represents the personal thinking of each particle. The cognitive component of a particle takes the best position found so far by this particle as the desired input to make the particle move toward its own best positions.

$c_2 r_{2,n}^{i,d}[\mathrm{Pg}_n^{i,d} - X_n^{i,d}]$ is known as the "social" component, which represents the collaborative behavior of the particles to find the global optimal solution. The social component always pulls the particles toward the best position found by its neighbors.

In a nutshell, a PSO algorithm has the following advantages: (1) It has versatility and does not rely on the problem information; (2) it has a memory capacity to retain local and global optimal information; (3) it is easy to implement. Given the versatility and effectiveness of PSO, it is widely used to solve practical problems such as artificial neural networks [23, 43], chemical systems [17], power systems [5, 6], mechanical design [28], communications [71], robotics [32,63], economy [45,47], image processing [46], bio-informatics [53,64], medicine [58], and industrial engineering [40,60]. Note that swarms have also been used in many engineering applications, for example, in collective robotics where there are teams of robots working together by communicating over a communication network; see [38] for a stability analysis and many related references.

To enable and to enhance further applications, much work has also been devoted to improving the PSO algorithms. Because the original model is similar to a mobile multi-agent system and each parameter describes a special character of natural swarm behavior, one can improve the performance of PSO according to the physical meanings of these parameters [39, 48, 52, 54, 72]. The first significant improvement was proposed by Shi and Eberhart in [59]. They suggested to add a new parameter w as an "inertia constant", which results in fast convergence. The modified equation of (2.1) is

$$\begin{aligned} v_{n+1}^{i,d} &= w v_n^{i,d} + c_1 r_{1,n}^{i,d}[\mathrm{Pr}_n^{i,d} - X_n^{i,d}] + c_2 r_{2,n}^{i,d}[\mathrm{Pg}_n^{i,d} - X_n^{i,d}], \\ X_{n+1}^{i,d} &= X_n^{i,d} + v_{n+1}^{i,d}. \end{aligned} \tag{2.3}$$

Another significant improvement was due to Clerc and Kennedy [16]. They introduced a constriction coefficient χ and then proposed to modify (2.1) as

$$\begin{aligned} v_{n+1}^{i,d} &= \chi(v_n^{i,d} + c_1 r_{1,n}^{i,d}[\mathrm{Pr}_n^{i,d} - X_n^{i,d}] + c_2 r_{2,n}^{i,d}[\mathrm{Pg}_n^{i,d} - X_n^{i,d}]), \\ X_{n+1}^{i,d} &= X_n^{i,d} + v_{n+1}^{i,d}. \end{aligned} \tag{2.4}$$

This constriction coefficient can control the "explosion" of the PSO and ensure the convergence. Some researchers also considered using "good" topologies of particle connection, in particular adaptive ones (e.g., [14, 42, 44]).

There has been much development on mathematical analysis for the convergence of PSO algorithms as well. Although most researchers prefer to use discrete system [13, 16, 62, 66], there are some works on continuous-time models [18, 41]. Some recent work such as [15, 19, 24, 37, 51, 65] provides guidelines for selecting PSO parameters leading to convergence, divergence, or oscillation of the swarm's particles. The aforementioned work also gives rise to several PSO variants. Nowadays, it is widely recognized that purely deterministic approach is inadequate in reflecting the exploration and exploitation aspects brought by stochastic variables. However, as criticized by Pedersen [50], the analysis is often oversimplified. For example, the swarm is often assumed to have only one particle; stochastic variables (namely, $r_{1,n}$, $r_{2,n}$) are not used; the points of attraction, i.e., the particle's best known position Pr and the swarm's best known position Pg, are normally assumed to remain constant throughout the optimization process.

In this chapter, we study convergence of PSO by using stochastic approximation methods. In the past, some authors have considered using stochastic approximation combined with PSO to enhance the performance or select parameters (e.g., [27]). But to the best of our

knowledge, the only chapter using stochastic approximation methods to analyze the dynamics of the PSO so far is by Chen and Li [15]. They designed a special PSO procedure and assumed that (i) Pr_n^i and Pg_n^i are always within a finite domain; (ii) with P^* representing the global optimal positions in the solution space, and $\|P^*\| < \infty$. $\lim_{n\to\infty} \mathrm{Pr}_n \to P^*$ and $\lim_{n\to\infty} \mathrm{Pg}_n \to P^*$. Using assumption (i), they proved the convergence of the algorithm in the sense of with probability one. With additional assumption (ii), they showed that the swarm will converge to P^*. Despite the interesting development, their assumptions (i) and (ii) appear to be rather strong. Moreover, they added some specific terms in the PSO procedure. So their algorithm is different from the traditional PSOs (2.1)-(2.4). In this chapter, we consider a general form of PSO algorithms. We introduce four coefficients ε, κ_1, κ_2, and χ and rewrite the PSOs in a stochastic approximation setup. Then we analyze its convergence using weak convergence method. We prove that a suitably interpolated sequence of swarms converge to the solution of an ordinary differential equation. Moreover, convergence rates are derived by using a sequence of centered and scaled estimation errors.

The rest of the chapter is arranged as follows. Section 2.2 presents the setup of our algorithm. Section 2.3 studies the convergence and Section 2.4 analyzes the rate of convergence. Section 2.5 proceeds with several numerical simulation examples to illustrate the convergence of our algorithms. Finally, Section 2.6 provides a few further remarks.

2.2 Formulation

First, we will introduce some notations used in this chapter. We use $|\cdot|$ to denote a Euclidean norm. A point θ in a Euclidean space is a column vector; the ith component of θ is denoted by θ^i; $\mathrm{diag}(\theta)$ is a diagonal matrix whose diagonal elements are the elements of θ; I denotes the identity matrix of appropriate dimension; z' denotes the transposition of z; the notation

$O(y)$ denotes a function of y satisfying $\sup_y |O(y)|/|y| < \infty$, and $o(y)$ denotes a function of y satisfying $|o(y)|/|y| \to 0$, as $y \to 0$. In particular, $O(1)$ denotes the boundedness and $o(1)$ indicates convergence to 0. Throughout the chapter, for convenience, we use K to denote a generic positive constant with the convention that the value of K may be different for different usage.

In this chapter, without loss of generality, we assume that each particle is a one-dimensional scalar. Note that each particle can be a multi-dimensional vector, which does not introduce essential difficulties in the analysis; only the notation is a bit more complex. We introduce four parameters ε, κ_1, κ_2, and χ. Suppose there are r particles, then the PSO algorithm can be expressed as

$$\begin{aligned}\begin{bmatrix} v_{n+1} \\ X_{n+1} \end{bmatrix} = \begin{bmatrix} v_n \\ X_n \end{bmatrix} + \varepsilon \Bigg(& \begin{bmatrix} \kappa_1 I & -\chi(c_1\mathrm{diag}(r_{1,n}) + c_2\mathrm{diag}(r_{2,n})) \\ \kappa_2 I & -\chi(c_1\mathrm{diag}(r_{1,n}) + c_2\mathrm{diag}(r_{2,n})) \end{bmatrix} \begin{bmatrix} v_n \\ X_n \end{bmatrix} \\ & + \chi \begin{bmatrix} c_1\mathrm{diag}(r_{1,n}) & c_2\mathrm{diag}(r_{2,n}) \\ c_1\mathrm{diag}(r_{1,n}) & c_2\mathrm{diag}(r_{2,n}) \end{bmatrix} \begin{bmatrix} \mathrm{Pr}(\theta_n, \eta_n) \\ \mathrm{Pg}(\theta_n, \eta_n) \end{bmatrix} \Bigg), \end{aligned} \tag{2.5}$$

where c_1 and c_2 represent the acceleration coefficients, $X_n = [X_n^1, \ldots, X_n^r]' \in \mathbb{R}^r$, $v_n = [v_n^1, \ldots, v_n^r]' \in \mathbb{R}^r$, $\theta_n = (X_n, v_n)'$, r_1, r_2 are r-dimensional random vectors in which each component is uniformly distributed in $(0, 1)$, and $\mathrm{Pr}(\theta, \eta)$ and $\mathrm{Pg}(\theta, \eta)$ are two non-linear functions depending on $\theta = (X, v)'$ as well as on a "noise" η, and $\varepsilon > 0$ is a small parameter representing the stepsize of the iterations.

Remark 2.1. Note that for a large variety of cases, the structures and the forms of $\mathrm{Pr}(\theta, \eta)$ and $\mathrm{Pg}(\theta, \eta)$ are not known. This is similar to the situation in a stochastic optimization problem in which the objective function is not known precisely. Thus, stochastic approximation methods are well suited. As it is well known that stochastic approximation methods are

very useful for treating optimization problems in which the form of the objective function is not known precisely, or too complex to compute. The beauty of such stochastic iteratively defined procedures is that one need not know the precise form of the functions.

If there is no noise term η_n, let $\varepsilon = 0.01$, $\chi = 72.9$, $\kappa_1 = -27.1$, and $\kappa_2 = 72.9$, then (2.5) is equivalent to (2.3) when $w = 0.729$ or (2.4) when $\chi = 0.729$. Thus (2.5) is a generalization of (2.1)-(2.4). So a lot of approaches of tuning parameters (e.g., [10, 49, 70]) could also be applied.

Remark 2.2. In the proposed algorithm, we use a constant stepsize. The stepsize $\varepsilon > 0$ is a small parameter. As is well recognized (see [11, 90]), constant stepsize algorithms have the ability to track slight time variation and is more preferable in many applications. In the convergence and rate of convergence analysis, we let $\varepsilon \to 0$. In the actual computation, ε is just a constant. It need not go to 0. This is the same as one carries out any computational problem in which the analysis requires the iteration number going to infinity. However, in the actual computing, one only executes the procedure finitely many steps.

In (2.5), r_1 and r_2 are used to reflect the exploration of particles. Rearranging terms of (2.5) and considering that $E[c_1\text{diag}(r_{1,n})] = 0.5c_1 I$ and $E[c_2\text{diag}(r_{2,n})] = 0.5c_2 I$, it can be

rewritten as

$$\begin{aligned}\begin{bmatrix} v_{n+1} \\ X_{n+1} \end{bmatrix} = \begin{bmatrix} v_n \\ X_n \end{bmatrix} + \varepsilon \Bigg\{ & \begin{bmatrix} \kappa_1 I & -0.5\chi(c_1+c_2)I \\ \kappa_2 I & -0.5\chi(c_1+c_2)I \end{bmatrix} \begin{bmatrix} v_n \\ X_n \end{bmatrix} \\ & + \chi \begin{bmatrix} 0.5c_1 I & 0.5c_2 I \\ 0.5c_1 I & 0.5c_2 I \end{bmatrix} \begin{bmatrix} \Pr(\theta_n, \eta_n) \\ \mathrm{Pg}(\theta_n, \eta_n) \end{bmatrix} \\ & + \chi \begin{bmatrix} 0 & -(c_1 \mathrm{diag}(r_{1,n}) + c_2 \mathrm{diag}(r_{2,n}) - 0.5c_1 I - 0.5c_2 I) \\ 0 & -(c_1 \mathrm{diag}(r_{1,n}) + c_2 \mathrm{diag}(r_{2,n}) - 0.5c_1 I - 0.5c_2 I) \end{bmatrix} \begin{bmatrix} v_n \\ X_n \end{bmatrix} \\ & + \chi \begin{bmatrix} c_1 \mathrm{diag}(r_{1,n}) - 0.5c_1 I & c_2 \mathrm{diag}(r_{2,n}) - 0.5c_2 I \\ c_1 \mathrm{diag}(r_{1,n}) - 0.5c_1 I & c_2 \mathrm{diag}(r_{2,n}) - 0.5c_2 I \end{bmatrix} \begin{bmatrix} \Pr(\theta_n, \eta_n) \\ \mathrm{Pg}(\theta_n, \eta_n) \end{bmatrix} \Bigg\}. \end{aligned} \tag{2.6}$$

Denote

$$\begin{aligned} & \theta_n = [v_n, X_n]' \in \mathbb{R}^{2r}, \\ & M = \begin{bmatrix} \kappa_1 I & -0.5\chi(c_1+c_2)I \\ \kappa_2 I & -0.5\chi(c_1+c_2)I \end{bmatrix}, \\ & P(\theta_n, \eta_n) = \chi \begin{bmatrix} 0.5c_1 I & 0.5c_2 I \\ 0.5c_1 I & 0.5c_2 I \end{bmatrix} \begin{bmatrix} \Pr(\theta_n, \eta_n) \\ \mathrm{Pg}(\theta_n, \eta_n) \end{bmatrix}, \end{aligned} \tag{2.7}$$

and $W(\theta_n, r_{1,n}, r_{2,n}, \eta_n)$ to be the sum of the last two terms in the curly braces of (2.6). Then (2.6) can be expressed as a stochastic approximation algorithm

$$\theta_{n+1} = \theta_n + \varepsilon[M\theta_n + P(\theta_n, \eta_n) + W(\theta_n, r_{1,n}, r_{2,n}, \eta_n)]. \tag{2.8}$$

One of the challenges in analyzing the convergence of PSO is that the concrete forms of $\Pr(\theta_n, \eta_n)$ and $\mathrm{Pg}(\theta_n, \eta_n)$ are unknown. However, this will not concern us. As mentioned before, stochastic approximation methods are known to have advantages in treating such

situations. We shall use the following assumptions.

(A1) The $\Pr(\cdot,\eta)$ and $\mathrm{Pg}(\cdot,\eta)$ are continuous for each η. For each bounded θ, $E|P(\theta,\eta_n)|^2 < \infty$ and $E|W(\theta,r_{1,n},r_{2,n},\eta_n)|^2 < \infty$. There exist continuous functions $\overline{\Pr}(\theta)$ and $\overline{\mathrm{Pg}}(\theta)$ such that

$$\begin{aligned} &\frac{1}{n}\sum_{j=m}^{n+m-1} E_m \Pr(\theta,\eta_j) \to \overline{\Pr}(\theta) \ \text{ in probability,} \\ &\frac{1}{n}\sum_{j=m}^{n+m-1} E_m \mathrm{Pg}(\theta,\eta_j) \to \overline{\mathrm{Pg}}(\theta) \ \text{ in probability,} \end{aligned} \tag{2.9}$$

where E_m denotes the conditional expectation on the σ-algebra $\mathcal{F}_m = \{\theta_0, r_{i,j}, i = 1,2, \eta_j : j < m\}$. Moreover, for each θ in a bounded set,

$$\begin{aligned} &\sum_{j=n}^{\infty} |E_n \Pr(\theta,\eta_j) - \overline{\Pr}(\theta)| < \infty, \\ &\sum_{j=n}^{\infty} |E_n \mathrm{Pg}(\theta,\eta_j) - \overline{\Pr}(\theta)| < \infty. \end{aligned} \tag{2.10}$$

(A2) Define

$$\overline{P}(\theta) = \chi \begin{bmatrix} 0.5c_1 I & -0.5c_2 I \\ 0.5c_1 I & -0.5c_2 I \end{bmatrix} \begin{bmatrix} \overline{\Pr}(\theta) \\ \overline{\mathrm{Pg}}(\theta) \end{bmatrix}.$$

The ordinary differential equation

$$\frac{d\theta(t)}{dt} = M\theta(t) + \overline{P}(\theta(t)) \tag{2.11}$$

has a unique solution for each initial condition $\theta(0) = (\theta_0^1, \ldots, \theta_0^{2r})'$.

(A3) For $i = 1,2$, $\{r_{i,n}\}$ and $\{\eta_n\}$ are mutually independent; $\{r_{i,n}\}$ are i.i.d. sequences of random variables with each component being uniformly distributed in $(0,1)$.

Remark 2.3. Condition (A1) is satisfied by a large class of functions and random variables. The continuity is assumed for convenience. In fact, only weak continuity is needed so we can in fact deal with indicator type of functions whose expectations are continuous.

In fact, (2.9) mainly requires that $\{\Pr(\theta, \eta_n)\}$ is a sequence that satisfies a law of large number type of condition, although it is weaker than the usual weak law of large numbers. Condition (2.10) is modeled by the mixing type condition. For instance, we may assume that for each bounded random vector θ and each $T < \infty$, either

$$\lim_{j\to\infty,\Delta\to 0} E \sup_{|Y|\le\Delta} |\Pr(\theta + Y, \eta_j) - \Pr(\theta, \eta_j)| = 0, \quad \text{or}$$
$$\lim_{n\to\infty,\Delta\to 0} \frac{1}{n} \sum_{j=m}^{m+n-1} E \sup_{|Y|\le\Delta} |\Pr(\theta + Y, \eta_j) - \Pr(\theta, \eta_j)| = 0.$$

Apparently, the second alternative is even weaker. With either of this assumption, all of the subsequent development follows, but the argument is more complex. Under the above condition, one can treat discontinuity involving sign function or indicator function among others. For the corresponding stochastic approximation algorithms, see [106, p. 100]; the setup in [90] is even more general, which allows in addition to the discontinuity, the functions involved to be time dependent. Inserting the conditional expectation is much weaker than without. For example, if $\{\eta_n\}$ is a sequence of i.i.d. random variables with distribution function F_η, then for each θ, $\overline{\Pr}(\theta) = E\Pr(\theta, \eta_1) = \int \Pr(\theta, \zeta) F_\eta(d\zeta)$, so (2.9) is easily verified. Likewise, if $\{\eta_n\}$ is a martingale difference sequence, the condition is also satisfied. Next, if $\{\eta_n\}$ is a moving average sequence driven by a martingale difference sequence, (2.9) is also satisfied. In addition, if $\{\eta_n\}$ is a mixing sequence [97, p.166] with the mixing rate decreasing to 0, the condition is also satisfied. Note that in a mixing sequence, there can be infinite correlations and the remote past and distant future are only asymptotically uncorrelated.

In the simplest additive noise case, i.e., $\Pr(\theta, \eta) = \Pr(\theta) + \eta$, then the condition is mainly on the noise sequence $\{\eta_n\}$. Condition (2.10) is modeled after the so-called mixing inequality; see [106, p.82] and references therein. Suppose that $\{\Pr(\theta, \eta_n)\}$ is a stationary mixing sequence with mean $\overline{\Pr}(\theta)$ and mixing rate ϕ_n such that $\sum_n \phi_n^{1/2} < \infty$, then (2.10) is satisfied.

With these assumptions, we proceed to analyze the convergence and rates of convergence of PSO algorithms with general form (2.8). The scheme is a constant-step-size stochastic approximation algorithm with step size ε. Our interest lies in obtaining convergence and rates of convergence as $\varepsilon \to 0$. We emphasize that in the actual computation, it is not necessary to modify it as the generalized PSO form (2.8). This generalized PSO form is simply a convenient form that allows us to analyze the algorithm by using methods of stochastic approximation.

2.3 Convergence

This section is devoted to obtaining asymptotic properties of algorithm (2.8). In relation to PSO the word "convergence" typically means one of two things, although it is often not clarified which definition is meant and sometimes they are mistakenly thought to be identical. (i) Convergence may refer to the swarm's best known position Pg approaching (converging to) the optimum of the problem, regardless of how the swarm behaves. (ii) Convergence may refer to a swarm collapse in which all particles have converged to a point in the search space, which may or may not be the optimum. Since the convergence may rely on structure of the cost function if we use the first definition of convergence, we use the second one as the definition of convergence in this study. Our first result concerns the property of the algorithm

as $\varepsilon \to 0$ through an appropriate continuous-time interpolation. We define

$$\theta^{\varepsilon}(t) = \theta_n \quad \text{for} \quad t \in [\varepsilon n, \varepsilon n + \varepsilon).$$

Then $\theta^{\varepsilon}(\cdot) \in D([0,T] : \mathbb{R}^{2r})$, which is the space of functions that are defined on $[0,T]$ taking values in $\mathbb{R}^{2r}$, and that are right continuous and have left limits endowed with the Skorohod topology [90, Chapter 7].

Theorem 2.4. *Under* (A1)-(A3), $\theta^{\varepsilon}(\cdot)$ *is tight in* $D([0,T] : \mathbb{R}^{2r})$. *Moreover, as* $\varepsilon \to 0$, $\theta^{\varepsilon}(\cdot)$ *converges weakly to* $\theta(\cdot)$, *which is a solution of* (2.11).

Remark 2.5. An equivalent way of stating the ODE limit (2.11) is to consider its associated martingale problem [106, pp. 15-16]. Consider the differential operator associated with $\theta(\cdot)$

$$\mathcal{L}f(\theta) = (\nabla f(\theta))'(M\theta + \overline{P}(\theta)),$$

and define

$$\widetilde{M}_f(t) = f(\theta(t)) - f(\theta(0)) - \int_0^t \mathcal{L}f(\theta(s))ds.$$

If $\widetilde{M}_f(\cdot)$ is a martingale for each $f(\cdot) \in C_0^1$ (C^1 function with compact support), then $\theta(\cdot)$ is said to solve a martingale problem with operator $\mathcal{L}$. Thus, an equivalent way to state the theorem is to prove that $\theta^{\varepsilon}(\cdot)$ converges weakly to $\theta(\cdot)$, which is a solution of the martingale problem with operator $\mathcal{L}$.

Proof of Theorem 2.4**.** To prove the tightness in $D([0,T] : \mathbb{R}^{2r})$, we first need to show

$$\lim_{K\to\infty} \limsup_{\varepsilon\to 0} P\{\sup_{t\le T} |\theta^{\varepsilon}(t)| \ge K\} = 0 \tag{2.12}$$

To avoid verifying (2.12), we define a process $\theta^{\varepsilon,N}(\cdot)$ satisfies $\theta^{\varepsilon,N}(t) = \theta^{\varepsilon}(t)$ up until the first exit from $S_N = \{x \in \mathbb{R}^{2r} : |x| \leq N\}$ and satisfies (2.12), the $\theta^{\varepsilon,N}(\cdot)$ is said to be an *N-truncation* of $\theta^{\varepsilon}(\cdot)$. Introduce a truncation function $q^N(\cdot)$ that is smooth and that satisfies $q^N(\theta) = 1$ for $|\theta| \leq N$, $q^N(\theta) = 0$ for $|\theta| \geq N+1$. Then the discrete system (2.8) is defined as

$$\theta_{n+1}^N = \theta_n^N + \varepsilon[M\theta_n^N + P(\theta_n^N, \eta_n) + W(\theta_n^N, r_{1,n}, r_{2,n}, \eta_n)]q^N(\theta_n^N), \tag{2.13}$$

using the N-truncation. Moreover, the N-truncated ODE and the operator $\mathcal{L}^N$ of the associated martingale problem can be defined as

$$\frac{d\theta^N(t)}{dt} = [M\theta^N(t) + \overline{P}(\theta^N(t))]q^N(\theta(t)), \tag{2.14}$$

and

$$\mathcal{L}^N f(\theta) = (\nabla f(\theta))'[M\theta + \overline{P}(\theta)]q^N(\theta), \tag{2.15}$$

respectively.

To prove the theorem, we proceed to verify the following claims: (a) for each N, $\{\theta^{\varepsilon,N}(\cdot)\}$ is tight. By virtue of the Prohorov theorem [90, p.229], we can extract a weakly convergent subsequence. For notational simplicity, we still denote the subsequence by $\{\theta^{\varepsilon,N}(\cdot)\}$ with limit denoted by $\theta^N(\cdot)$.

(b) $\theta^N(\cdot)$ is a solution of the martingale problem with operator $\mathcal{L}^N$.

Using the uniqueness of the limit, passing to the limit as $N \to \infty$, and by the corollary in [106, p.44], $\{\theta^{\varepsilon}(\cdot)\}$ converges weakly to $\theta(\cdot)$.

Now we start to prove claims (a) and (b).

(a) Tightness. For any $\delta > 0$, let $t > 0$ and $s > 0$ such that $s \leq \delta$, and t, $t + \delta \in [0, T]$.

Note that

$$\theta^{\varepsilon,N}(t+s)-\theta^{\varepsilon,N}(t)=\varepsilon\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}(M\theta_k^N+P(\theta_k^N,\eta_k)+W(\theta_k^N,r_{1,k},r_{2,k},\eta_k))q^N(\theta_k^N).$$

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

Using the Cauchy-Schwarz inequality,

$$\varepsilon^2E_t^\varepsilon\left|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}M\theta_k^Nq^N(\theta_k^N)\right|^2\le\varepsilon Ks\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}E_t^\varepsilon\left|\theta_k^Nq^N(\theta_k^N)\right|^2. \tag{2.16}$$

where E_t^ε denotes the expectation conditioned on the σ-algebra $\mathcal{F}_t^\varepsilon$. Likewise,

$$\varepsilon^2E_t^\varepsilon\left|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}W(\theta_k^N,r_{1,k},r_{2,k},\eta_k)q^N(\theta_k^N)\right|^2\le Ks^2, \tag{2.17}$$

and

$$\varepsilon^2E_t^\varepsilon\left|\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}P(\theta_k^N,\eta_k)q^N(\theta_k^N)\right|^2\le Ks^2. \tag{2.18}$$

So we have

$$E_t^\varepsilon\left|\theta^{\varepsilon,N}(t+s)-\theta^{\varepsilon,N}(t)\right|^2\le K\varepsilon s\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}\sup_{t/\varepsilon\le k\le(t+s)/\varepsilon-1}E_t^\varepsilon|\theta_k^Nq^N(\theta_k^N)|^2+Ks^2. \tag{2.19}$$

As a result, there is a $\varsigma^\varepsilon(\delta)$ such that

$$E_t^\varepsilon|\theta^{\varepsilon,N}(t+s)-\theta^{\varepsilon,N}(t)|^2\le E_t^\varepsilon\varsigma^\varepsilon(\delta)\ \text{ for all }\ 0\le s\le\delta,$$

and that $\lim_{\delta\to 0}\limsup_{\varepsilon\to 0} E\varsigma^{\varepsilon}(\delta) = 0$. The tightness of $\{\theta^{\varepsilon,N}(\cdot)\}$ then follows from [106, p.47].

(b) Characterization of the limit. To characterize the limit process, we need to work with a continuously differentiable function with compact support $f(\cdot)$. Choose m_ε so that $m_\varepsilon \to \infty$ as $\varepsilon \to 0$ but $\delta_\varepsilon = \varepsilon\, m_\varepsilon \to 0$. Using the recursion (2.13),

$$
\begin{aligned}
f(\theta^{\varepsilon,N}(t+s)) - f(\theta^{\varepsilon,N}(t)) &= \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} [f(\theta^N_{lm_\varepsilon+m_\varepsilon}) - f(\theta^N_{lm_\varepsilon})] \\
&= \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} [M\theta^N_k + \overline{P}(\theta^N_k)] q^N(\theta^N_k) \\
&\quad + \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} [P(\theta^N_k, \eta_k) - \overline{P}(\theta^N_k)] q^N(\theta^N_k) \\
&\quad + \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} W(\theta^N_k, r_{1,k}, r_{2,k}, \eta_k) q^N(\theta^N_k) \\
&\quad + \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \Bigg\{ (\nabla f(\theta^{N+}_{lm_\varepsilon}) - \nabla f(\theta^N_{lm_\varepsilon}))' \\
&\qquad \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} [M\theta^N_k + P(\theta^N_k, \eta_k) \\
&\qquad\qquad + W(\theta^N_k, r_{1,k}, r_{2,k}, \eta_k)] q^N(\theta^N_k) \Bigg\},
\end{aligned}
\tag{2.20}
$$

where $\theta^{N+}_{lm_\varepsilon}$ is a point on the line segment joining $\theta^N_{lm_\varepsilon}$ and $\theta^N_{lm_\varepsilon+m_\varepsilon}$.

Our focus here is to characterize the limit. By the Skorohod representation [90, p.230], with a slight abuse of notation, we may assume that $\theta^{\varepsilon,N}(\cdot)$ converges to $\theta^N(\cdot)$ with probability one and the convergence is uniform on any bounded time interval. To show that $\{\theta^{\varepsilon,N}(\cdot)\}$ is a solution of the martingale problem with operator $\mathcal{L}^N$, it suffices to show that for any

$f(\cdot) \in C_0^1$, the class of functions that are continuously differentiable with compact support,

$$\widetilde{M}_f^N(t) = f(\theta^N(t)) - f(\theta^N(0)) - \int_0^t \mathcal{L}^N f(\theta^N(u))du$$

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 $t_i \leq t$ with $i \leq \kappa$,

$$\begin{aligned} &Eh(\theta^N(t_i) : i \leq \kappa)[\widetilde{M}_f^N(t+s) - \widetilde{M}_f^N(t)] \\ &= Eh(\theta^N(t_i) : i \leq \kappa) \times [f(\theta^N(t+s)) - f(\theta^N(t)) - \int_t^{t+s} \mathcal{L}^N f(\theta^N(u))du] \\ &= 0. \end{aligned} \tag{2.21}$$

To verify (2.21), we begin with the process indexed by ε. For notational simplicity, denote

$$\widetilde{h} = h(\theta^N(t_i) : i \leq \kappa), \quad \widetilde{h}^\varepsilon = h(\theta^{\varepsilon,N}(t_i) : i \leq \kappa). \tag{2.22}$$

Then the weak convergence and the Skorohod representation together with the boundedness and the continuity of $f(\cdot)$ and $h(\cdot)$ yield that as $\varepsilon \to 0$,

$$E\widetilde{h}^\varepsilon[f(\theta^{\varepsilon,N}(t+s)) - f(\theta^{\varepsilon,N}(t)) \to E\widetilde{h}[f(\theta^N(t+s)) - f(\theta^N(t))].$$

For the last term of (2.20), as $\varepsilon \to 0$, since $f(\cdot) \in C_0^1$,

$$\begin{aligned} E\widetilde{h}^\varepsilon \varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \Bigg\{ &(\nabla f(\theta_{lm_\varepsilon}^{N+}) - \nabla f(\theta_{lm_\varepsilon}^N))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} [M\theta_k^N + P(\theta_k^N, \eta_k) \\ &+ W(\theta_k^N, r_{1,k}, r_{2,k}, \eta_k)]q^N(\theta_k^N) \Bigg\} \\ &= O(\varepsilon) \to 0. \end{aligned} \tag{2.23}$$

For the next to the last term of (2.20),

$$\begin{aligned}&\lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} W(\theta^N_k, r_{1,k}, r_{2,k}, \eta_k) q^N(\theta^N_k)\Big] \\ &\quad = \lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\Big[\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \frac{\delta_\varepsilon}{m_\varepsilon} \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E_{lm_\varepsilon} W(\theta^N_k, r_{1,k}, r_{2,k}, \eta_k) q^N(\theta^N_k)\Big].\end{aligned} \tag{2.24}$$

Using (A1) and (A3),

$$\frac{1}{m_\varepsilon} \sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} E_{lm_\varepsilon} W(\theta^N_{lm_\varepsilon}, r_{1,j}, r_{2,j}, \eta_j) q^N(\theta^N_{lm_\varepsilon}) \to 0$$

in probability, we obtain that

$$E\widetilde{h}^{\varepsilon}\Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} W(\theta^N_k, r_{1,k}, r_{2,k}, \eta_k) q^N(\theta^N_k)\Big] \to 0. \tag{2.25}$$

Using (A1), we obtain

$$E\widetilde{h}^{\varepsilon}\Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (P(\theta^N_k, \eta_k) - \overline{P}(\theta^N_k)) q^N(\theta^N_k)\Big] \to 0. \tag{2.26}$$

Next, we consider the first term. We have

$$\begin{aligned}&\lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (M\theta^N_k + \overline{P}(\theta^N_k)) q^N(\theta^N_k)\Big] \\ &\quad = \lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} (\nabla f(\theta^N_{lm_\varepsilon}))' \times \sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} (M\theta^N_{lm_\varepsilon} + \overline{P}(\theta^N_{lm_\varepsilon})) q^N(\theta^N_{lm_\varepsilon})\Big].\end{aligned} \tag{2.27}$$

Thus, to get the desired limit, we need only examine the last two lines above. Let $\varepsilon\, lm_\varepsilon \to u$ as $\varepsilon \to 0$. Then for all k satisfying $lm_\varepsilon \le k \le lm_\varepsilon + m_\varepsilon - 1$, $\varepsilon\, k \to u$ since $\delta_\varepsilon \to 0$. As a

result,

$$\begin{aligned}&\lim_{\varepsilon\to 0} E\widetilde{h}^{\varepsilon}\Big[\varepsilon\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon}(\nabla f(\theta^N_{lm_\varepsilon}))'\times\sum_{k=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}(M\theta^N_{lm_\varepsilon}+\overline{P}(\theta^N_{lm_\varepsilon}))q^N(\theta^N_{lm_\varepsilon})\Big]\\&=E\widetilde{h}\Big[\int_t^{t+s}(\nabla f(\theta^N(u)))'(M(\theta^N(u))+\overline{P}(\theta(u)))q^N(\theta(u))du\Big].\end{aligned}\tag{2.28}$$

The desired result then follows. □

To proceed, consider (2.11). For simplicity, suppose that there is a unique stationary point θ^*. Denote $\overline{\mathrm{Pr}}(\theta^*)=\mathrm{Pr}^*$ and $\overline{\mathrm{Pg}}(\theta^*)=\mathrm{Pg}^*$. By the inversion formula of partitioned matrix [61], solving $M\theta^*+\overline{P}(\theta^*)=0$ yields that the equilibrium point of the ODE satisfies

$$\begin{aligned}\theta^*&=\begin{bmatrix}\kappa_1 I & -0.5\chi(c_1+c_2)I\\ \kappa_2 I & -0.5\chi(c_1+c_2)I\end{bmatrix}^{-1}\times\begin{bmatrix}-0.5\chi(c_1\mathrm{Pr}^*+c_2\mathrm{Pg}^*)\\ -0.5\chi(c_1\mathrm{Pr}^*+c_2\mathrm{Pg}^*)\end{bmatrix}\\&=\begin{bmatrix}0\\ \frac{c_1\,\mathrm{Pr}^*+c_2\mathrm{Pg}^*}{c_1+c_2}\end{bmatrix}.\end{aligned}\tag{2.29}$$

Corollary 2.6. *Suppose that the stationary point θ^* is asymptotically stable in the sense of Lyapunov and that $\{\theta_n\}$ is tight. Then for any $t_\varepsilon\to\infty$ as $\varepsilon\to 0$, $\theta^\varepsilon(\cdot+t_\varepsilon)$ converges weakly to θ^*.*

Proof. Define $\widetilde{\theta}^\varepsilon(\cdot)=\theta^\varepsilon(\cdot+t_\varepsilon)$. Let $T>0$ and consider the pair $\{\widetilde{\theta}^\varepsilon(\cdot),\widetilde{\theta}^\varepsilon(\cdot-T)\}$. Using the same argument as in the proof of Theorem 2.4, $\{\widetilde{\theta}^\varepsilon(\cdot),\ \widetilde{\theta}^\varepsilon(\cdot-T)\}$ is tight. Select a convergent subsequence with limit denoted by $(\theta(\cdot),\theta_T(\cdot))$. Then $\theta(0)=\theta_T(T)$. The value of $\theta_T(0)$ is not known, but all such $\theta_T(0)$, over all T and convergent subsequences, belong to a tight set. This together with the stability and Theorem 2.4 implies that for any $\Delta>0$, there is a T_Δ such that for $T>T_\Delta$, $P(\theta_T(T)\in U_\Delta(\theta^*))>1-\Delta$, where $U_\Delta(\theta^*)$ is a

Δ-neighborhood of θ^*. The desired result then follows. □

In Corollary 2.6, we used the tightness of the set $\{\theta_n\}$, which can be proved using the argument of Lemma 2.9. The result indicates that as the stepsize $\varepsilon \to 0$ and $n \to \infty$ with $n\varepsilon \to \infty$, θ_n converges to θ^* in the sense in probability. Note that if θ^* turns out to the optimum of the search space, then θ_n converges to the optimum.

Remark 2.7. Note that for notational simplicity, we have assumed that there is a unique stationary point of (2.11). As far as the convergence is concerned, one need not assume that there is only one θ^*. See how multimodal cases can be handled in the related stochastic approximation problems in [90, Chpaters 5, 6, 8]. In fact, for the multimodal cases, we can show that $\theta^\varepsilon(\cdot + t_\varepsilon)$ converges in an appropriate sense to the set of the stationary points. Thus Corollary 2.6 can be modified. In the rate of convergence study, [25] suggested an approach using conditional distribution, which is a modification of a single stationary point. If multiple stationary points are involved, we can simply use the approach of [25] combined with our weak convergence analysis. The notation will be a bit more complex, but main idea still rest upon the basic analysis method to be presented in the next section. It seems to be more instructive to present the main ideas, so we choose the current setting.

2.4 Rate of Convergence

Once the convergence of a stochastic approximation algorithm is established, the next task is to ascertain the convergence rate. To study the convergence rate, we take a suitably scaled sequence$z_n = (\theta_n - \theta^*)/\varepsilon^\alpha$, for some $\alpha > 0$. The idea is to choose α such that z_n converges (in distribution) to a nontrivial limit. The scaling factor α together with the asymptotic covariance of the scaled sequence gives us the rate of convergence. That is, the scaling tells us the dependence of the estimation error $\theta_n - \theta^*$ on the step size, and the asymptotic

covariance is a mean of assessing "goodness" of the approximation. Here the factor $\alpha = 1/2$ is used. To some extent, this is dictated by the well-known central limit theorem. For related work on convergence rate of various stochastic approximation algorithms, see [31, 67].

As mentioned above, by using the definition of the rate of convergence, we are effectively dealing with convergence in the distributional sense. In lieu of examining the discrete iteration directly, we are again taking continuous-time interpolations. Three assumptions are provided in what follows.

(A4) The following conditions hold:

(i) in a neighborhood of θ^*, $\mathrm{Pr}(\cdot, \eta)$ and $\mathrm{Pg}(\cdot, \eta)$ are continuously differentiable for each η, and the second derivatives (w.r.t. θ) of $W(\cdot, r_1, r_2, \eta)$ and $P(\cdot, \eta)$ exist and are continuous.

(ii) denoting by E_m the conditional expectation on the σ-algebra $\mathcal{F}_m = \{\theta_0,\ r_{1j}, r_{2j}, \eta_j : j < m\}$, and by ζ_θ the first partial derivative w.r.t. θ of $\zeta = W$ or P, resp., for each positive integer m, as $n \to \infty$,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=m}^{n+m-1} E_m \mathrm{Pr}_\theta(\theta, \eta_j) \to \overline{\mathrm{Pr}}_\theta(\theta) \ \text{ in probability,} \\
&\frac{1}{n}\sum_{j=m}^{n+m-1} E_m \mathrm{Pg}_\theta(\theta, \eta_j) \to \overline{\mathrm{Pg}}_\theta(\theta) \ \text{ in probability,} \\
&\sum_{j=m}^{\infty} |E_m W_\theta(\theta^*, r_{1,j}, r_{2,j}, \eta_j)| < \infty, \\
&\sum_{j=m}^{\infty} |E_m P_\theta(\theta^*, \eta_j) - \overline{P}_\theta(\theta^*)| < \infty.
\end{aligned} \tag{2.30}$$

(iii) The matrix $M + \overline{P}_\theta(\theta^*)$ is stable in that all of its eigenvalues are on the left half of the complex plane.

(iv) There is a twice continuously differentiable Lyapunov function $V(\cdot) : \mathbb{R}^{2r} \to \mathbb{R}$ such that

- $V(\theta) \to \infty$ as $|\theta| \to \infty$, and $V_{\theta\theta}(\cdot)$ is uniformly bounded.
- $|V_\theta(\theta)| \leq K(1 + V^{1/2}(\theta))$.
- $|M\theta + \overline{P}(\theta)|^2 \leq K(1 + V(\theta))$ for each θ.
- $V'_\theta(\theta)(M\theta + \overline{P}(\theta)) \leq -\lambda V(\theta)$ for some $\lambda > 0$ and each $\theta \neq \theta^*$.

(A5) $\sum_{j=m}^{\infty} |E\widetilde{W}'(\theta_m, r_{1,m}, r_{2,m}, \eta_m)\widetilde{W}(\theta_j, r_{1,j}, r_{2,j}, \eta_j)| < \infty$, where $\widetilde{W}(\theta, r_1, r_2, \eta) = P(\theta, \eta) - \overline{P}(\theta) + W(\theta, r_1, r_2, \eta)$.

(A6) The sequence $B^\varepsilon(t) = \sqrt{\varepsilon} \sum_{j=0}^{t/\varepsilon - 1} \widetilde{W}(\theta^*, r_{1,j}, r_{2,j}, \eta_j)$ converges weakly to $B(\cdot)$, a Brownian motion whose covariance Σt with $\Sigma \in \mathbb{R}^{2r \times 2r}$ given by

$$\begin{aligned}\Sigma = & E\widetilde{W}(\theta^*, r_{1,0}, r_{2,0}, \eta_0)\widetilde{W}'(\theta^*, r_{1,0}, r_{2,0}, \eta_0) \\ & + \sum_{k=1}^{\infty} E\widetilde{W}(\theta^*, r_{1,0}, r_{2,0}, \eta_0)\widetilde{W}'(\theta^*, r_{1,k}, r_{2,k}, \eta_k) \\ & + \sum_{k=1}^{\infty} E\widetilde{W}(\theta^*, r_{1,k}, r_{2,k}, \eta_k)\widetilde{W}'(\theta^*, r_{1,0}, r_{2,0}, \eta_0).\end{aligned} \tag{2.31}$$

Remark 2.8. Note that (A4)(ii) is another noise condition. The motivation is similar to Remark 2.3. The main difference of (2.9) and (2.10) and (2.30) is that (2.30) is on the derivative of the functions evaluated at the point θ^*. In fact, we only need the derivative exists in a neighborhood of this point only. This is because that we are analyzing the asymptotic normality locally. In view of this condition and condition of $\{r_{i,n}\}$,

$$\begin{aligned}&\frac{1}{n}\sum_{j=m}^{m+n-1} E_m W_\theta(\theta^*, r_{1,j}, r_{2,j}, \eta_j) \to 0 \text{ in probability,} \\ &\frac{1}{n}\sum_{j=m}^{m+n-1} E_m P_\theta(\theta^*, \eta_j) \to \overline{P}_\theta(\theta^*) \text{ in probability.}\end{aligned}$$

The traditional PSO algorithms do not allow non-additive noise, here we are treating a more general problem. Nonadditive noise can be allowed.

(A4)(iv) assumes the existence of a Lyapunov function. Only the existence is needed; its precise form need not be known. For simplicity, we have assumed the convergence of the scaled sequence to a Brownian motion in (A6); sufficient conditions are well known; see for example, [90, Section 7.4]. Before proceeding further, we first obtain a moment bound of θ_n.

Lemma 2.9. *Assume that* (A1)-(A6) *hold. Then there is an* N_ε *such that for all* $n > N_\varepsilon$, $EV(\theta_n) = O(\varepsilon)$.

Proof. To begin, it can be seen that

$$\begin{aligned} E_n V(\theta_{n+1}) - V(\theta_n) \leq \; & -\varepsilon\lambda\, V(\theta_n) + \varepsilon E_n V_\theta'(\theta_n)\widetilde{W}(\theta_n, r_{1,n}, r_{2,n}, \eta_n) \\ & + O(\varepsilon^2)(1 + V(\theta_n)), \end{aligned} \tag{2.32}$$

where θ_n^+ is on the line segment joining θ_n and θ_{n+1}. The bound in (2.32) follows from the growth condition in (A4)(iv), the last inequality follows from (A1). To proceed, we use the methods of perturbed Lyapunov functions, which entitles to introduce small perturbations to a Lyapunov function in order to make desired cancelation. Define a perturbation

$$V_1^\varepsilon(\theta, n) = \varepsilon \sum_{j=n}^{\infty} E_n V_\theta'(\theta)\widetilde{W}(\theta, r_{1,j}, r_{2,j}, \eta_j).$$

Note that

$$|V_1^\varepsilon(\theta, n)| = K\ \varepsilon(1 + V(\theta)). \tag{2.33}$$

Moreover,

$$\begin{aligned} & E_n V_1^\varepsilon(\theta_{n+1}, n+1) - V_1^\varepsilon(\theta_n, n) \\ & = O(\varepsilon^2)(V(\theta_n) + 1) - \varepsilon\ E_n V_\theta'(\theta_n)\widetilde{W}(\theta_n, r_{1,n}, r_{2,n}, \eta_n). \end{aligned} \tag{2.34}$$

Define $V^\varepsilon(\theta, n) = V(\theta) + V_1^\varepsilon(\theta, n)$.Using (2.32) and (2.34), we obtain

$$E_n V^\varepsilon(\theta_{n+1}, n+1) \leq\ (1 - \varepsilon\lambda) V^\varepsilon(\theta_n, n) + O(\varepsilon^2)(1 + V^\varepsilon(\theta_n, n)). \tag{2.35}$$

Choosing N_ε to be a positive integer such that $(1 - (\lambda\varepsilon/2))^{N_\varepsilon} \leq\ K\varepsilon$. Iterating on the recursion (2.35), taking expectation, and using the order of magnitude estimate (2.33), we

can then obtain

$$\begin{aligned} & E\, V^\varepsilon(\theta_{n+1}, n+1) \\ \leq\ & (1-\varepsilon\lambda)EV^\varepsilon(\theta_n, n) + O(\varepsilon^2)(1+V^\varepsilon(\theta_n, n)) \\ \leq\ & (1-\frac{\varepsilon\lambda}{2})^n EV^\varepsilon(\theta_0, 0) + O(\varepsilon) = O(\varepsilon). \end{aligned} \tag{2.36}$$

when $n > N_\varepsilon$. The second line of (2.36) follows from $1-\lambda\varepsilon+O(\varepsilon^2) \leq\ 1-\frac{\lambda\varepsilon}{2}$ for sufficiently small ε. Now using (2.33) again, we also have $EV(\theta_{n+1}) = O(\varepsilon)$. Thus the desired estimate follows. □

Define $z_n = (\theta_n - \theta^*)/\sqrt{\varepsilon}$. Then it is readily verified that

$$\begin{aligned} z_{n+1} = z_n\ & +\varepsilon(M + \overline{P}_\theta(\theta^*))z_n \\ & +\sqrt{\varepsilon}(P(\theta^*, \eta_n) - \overline{P}(\theta^*) + W(\theta^*, r_{1,n}, r_{2,n}, \eta_n)) \\ & +\varepsilon(P_\theta(\theta^*, \eta_n) - \overline{P}_\theta(\theta^*) \\ & \qquad +W_\theta(\theta^*, r_{1,n}, r_{2,n}, \eta_n))z_n + o(|z_n|^2). \end{aligned} \tag{2.37}$$

Corollary 2.10. *Assume that* (A1)-(A6) *hold. If the Lyapunov function is locally quadratic, i.e.,*

$$V(\theta) = (\theta-\theta^*)'Q(\theta-\theta^*) + o(|\theta-\theta^*|^2).$$

Then $EV(z_n) = O(1)$ *for all* $n > N_\varepsilon$.

Now we are in a position to study the asymptotic properties through weak convergence of appropriately interpolated sequence of z_n. Define $z^\varepsilon(t) = z_n$ for $t \in [(n-N_\varepsilon)\varepsilon, (n-N_\varepsilon)\varepsilon+\varepsilon]$. We can introduce a truncation sequence. That is, in lieu of $z^\varepsilon(\cdot)$, we let N be a fixed but otherwise arbitrary large positive integer and define $z^{\varepsilon,N}(\cdot)$ as an N-truncation of $z^\varepsilon(\cdot)$. That is, it is equal to $z^\varepsilon(\cdot)$ up until the first exit of the process from the sphere $S_N = \{|z| :\ |z| \leq\ N\}$ with radius N. Also define a truncation function $q^N(z) = 1$ if $z\ \in\ S_N, =\ 0$ if $z\ \in\ \mathbb{R}^r\ - S_{N+1}$, and is smooth. Corresponding to such a truncation, we also have a modified

operator with truncation (i.e., the functions used in the operator are all modified by use of $q^N(z)$). Then we proceed to establish the convergence of $z^{\varepsilon,N}(\cdot)$ as a solution of a martingale problem with the truncated operator. Then finally, letting $N \to \infty$, we use the uniqueness of the martingale problem to conclude the proof. The argument is similar to that of Section 2.3. For further technical details, we refer the reader to [90, pp. 284-285]. Such a truncation device is also widely used in the analysis of partial differential equations. For notational simplicity, we choose to simply assume the boundedness rather than go with the truncation route. Thus merely for notational simplicity, we suppose $z^{\varepsilon}(t)$ is bounded. For the rate of convergence, our focus is on the convergence of the sequence $z^{\varepsilon}(\cdot)$. We shall show that it converges to a diffusion process whose covariance matrix together with the scaling factor will provide us with the desired convergence rates. Although more complex than Theorem 2.4, we still use the martingale problem setup. To keep the presentation relatively brief, we shall only outline the main steps needed.

For any $t, s > 0$,

$$\begin{aligned} z^{\varepsilon}(t+s) - z^{\varepsilon}(t) = \varepsilon & \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} (M + \overline{P}_{\theta}(\theta^*))z_j \\ & + \sqrt{\varepsilon} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon} \widetilde{W}(\theta^*, r_{1,j}, r_{2,j}, \eta_j) \\ & + \varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon} \widetilde{W}_{\theta}(\theta^*, r_{1,j}, r_{2,j}, \eta_j) z_j. \end{aligned} \tag{2.38}$$

Note that for any $\delta > 0$, $t, s > 0$ with $s < \delta$,

$$E_t^{\varepsilon} \left| \sqrt{\varepsilon} \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \widetilde{W}(\theta^*, r_{1,j}, r_{2,j}, \eta_j) \right|^2 \leq K\varepsilon \left(\frac{t+s}{\varepsilon} - \frac{t}{\varepsilon} \right) = Ks \leq K\delta.$$

and similarly,

$$E_t^\varepsilon \left| \varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1} \widetilde{W}_\theta(\theta^*, r_{1,j}, r_{2,j}, \eta_j) z_j \right|^2 \leq Ks \leq K\delta.$$

Using Corollary 2.10 and similar argument as that of Theorem 2.4, we have the following result.

Lemma 2.11. *Assume conditions of Corollary* 2.10, $\{z^\varepsilon(\cdot)\}$ *is tight on* $D([0,T] : \mathbb{R}^{2r})$.

Next we can extract a convergent subsequence of $\{z^\varepsilon(\cdot)\}$. Without loss of generality, still denote the subsequence by $z^\varepsilon(\cdot)$ with limit $z(\cdot)$. For any $t, s > 0$, (2.38) holds. The way to derive the limit is similar to that of Theorem 2.4 using martingale problem formulation although the analysis is more involved. We proceed to show that the limit is the unique solution for the martingale problem with operator

$$Lf(z) = \frac{1}{2}\mathrm{tr}(\Sigma f_{zz}(z)) + (\nabla f(z))'(M + \overline{P}(\theta_*)), \tag{2.39}$$

for $f \in C_0^2$, C^2 functions with compact support.

Using similar notation as that of Section 2.3, redefine

$$\widetilde{h} = h(z(t_i) : i \leq \kappa), \quad \widetilde{h}^\varepsilon = h(z^\varepsilon(t_i) : i \leq \kappa). \tag{2.40}$$

By (A4) (ii), as $\varepsilon \to 0$

$$\begin{aligned} &E\widetilde{h}^\varepsilon \Big[\varepsilon \sum_{j=t/\varepsilon}^{(t+s)/\varepsilon} \widetilde{W}(\theta^*, r_{1,j}, r_{2,j}, \eta_j) z_j \Big] \\ &= E\widetilde{h}^\varepsilon \Big[\varepsilon \sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon} \sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1} \widetilde{W}(\theta^*, r_{1,j}, r_{2,j}, \eta_j) z_j \Big] \to 0. \end{aligned}$$

Using the notation as in Section 2.3,

$$E\widetilde{h}^{\varepsilon}\Big[\sum_{l=t/\delta_\varepsilon}^{(t+s)/\delta_\varepsilon}\frac{\delta_\varepsilon}{m_\varepsilon}\sum_{j=lm_\varepsilon}^{lm_\varepsilon+m_\varepsilon-1}\widetilde{W}(\theta^*,r_{1,j},r_{2,j},\eta_j)[z_j-z_{lm_\varepsilon}]\Big]$$
$$\to 0 \quad \text{as} \quad \varepsilon\to 0.$$

Moreover, by (A6) we have

$$\sqrt{\varepsilon}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon}\widetilde{W}(\theta^*,r_{1,j},r_{2,j},\eta_j)\to \int_t^{t+s} d\,B(u)$$

as $\varepsilon\to\ 0$. For the first term of (2.38), we have

$$E\widetilde{h}^{\varepsilon}\Big[\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}(M+\overline{P}_\theta(\theta^*))z_j\Big]$$
$$\to E\widetilde{h}\Big[\int_t^{t+s}(M+\overline{P}_\theta(\theta^*))z(u)du\Big]$$

as $\varepsilon\to\ 0$. Putting the aforementioned arguments together, we have the following theorem.

Theorem 2.12. *Under conditions* (A1)-(A7), $\{z^\varepsilon(\cdot)\}$ *converges to* $z(\cdot)$ *such that* $z(\cdot)$ *is a solution of the following stochastic differential equation*

$$dz=[M+\overline{P}_\theta(\theta^*)]zdt+\Sigma^{1/2}d\widehat{B}(t), \tag{2.41}$$

where $\widehat{B}(\cdot)$ *is a standard Brownian motion.*

Remark 2.13. To see what kind of functions and the associated ODE and SDE we are working with, we look at two simple examples. In the first example we use $F(x)=x^2$, take 2 particles, $\chi=1$, $\kappa_1=-0.271$, $\kappa_2=1$, $c_1=c_2=1.5$, and assume $\{\eta_k\}$ is an i.i.d. sequence

with mean $[0,0,0,0]'$ and variance I. Then

$$M = \begin{bmatrix} -0.271 & 0 & -1.5 & 0 \\ 0 & -0.271 & 0 & -1.5 \\ 1 & 0 & -1.5 & 0 \\ 0 & 1 & 0 & -1.5 \end{bmatrix}, \tag{2.42}$$

and the limit ODE is given by

$$\dot{\theta}(t) = M\theta(t).$$

Thus $\theta^* = [0,0,0,0]'$ is the minimizer of the swarm, and $P_\theta\ (\theta^*) = 0 \in \mathbb{R}^{4\times 4}$ (a 4×4 matrix with all entries being 0). In the standard optimization algorithm, one processor is running to approximate the optimum. Here, we have two particles running simultaneously. Note that θ has four components. Two of them represent the particles' positions, and the other two are the particles' speeds. At the end, both of the particles reach the minimum, representing something that might be called "overlapping." In addition, eventually the speeds of both particles reach 0 (or at resting point). As far as the rate of convergence is concerned, we conclude that $\theta_n - \theta^*$ decades in the order of $\sqrt{\varepsilon}$ (in the sense of convergence in distribution). Not only is the mean squares error of $(\theta_n - \theta^*)$ of the order ε, but also the interpolation of the scaled sequence z_n has a limit represented by a stochastic differential equation

$$dz = Mzdt + d\widehat{B}(t).$$

That is, (2.41) is satisfied with $P_\theta(\theta^*) = 0$ and $\Sigma = I$. As illustrated in [90], the scaling factor $\sqrt{\varepsilon}$ together with stationary covariance of the SDE gives us the rate of convergence. In terms of the swarm, loosely, we have $\theta_n - \theta^* \sim N(0,\ \varepsilon\ \Xi_0)$ [that is, $(\theta_n - \theta^*)$ is asymptot-

ically normal with mean $0 \in \mathbb{R}^4$ and covariance matrix $\varepsilon\, \Xi_0$], where Ξ_0 is the asymptotic covariance matrix that is the solution of the Lyapunov equation $M\, \Xi_0 + \Xi_0\, M' = -I$.

Likewise, in the second example, $F(x) = \sin x$ with $x \in [0, 1]$. We still take 2 particles, same parameters setting, and assume $\{\eta_k\}$ is the same i.i.d. sequence as before. Then M is as in (2.42), and

$$P_\theta(\theta^*) = \begin{bmatrix} 0.75 & 0 & 0.75 & 0 \\ 0 & 0.75 & 0 & 0.75 \\ 0.75 & 0 & 0.75 & 0 \\ 0 & 0.75 & 0 & 0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

It follows that (2.41) holds with

$$M + \overline{P}_\theta(\theta^*) = \begin{bmatrix} 1.229 & 0 & -1.5 & 0 \\ 0 & 1.229 & 0 & -1.5 \\ 2.5 & 0 & -1.5 & 0 \\ 0 & 2.5 & 0 & -1.5 \end{bmatrix}$$

and $\Sigma = I$. Similar to the previous example, we have that $\theta_n - \theta^*$ is asymptotically normal with mean 0 and covariance $\varepsilon\widetilde{\Xi}$, where $\widetilde{\Xi}$ is the asymptotic covariance satisfying the Lyapunov equation $(M + \overline{P}_\theta(\theta^*))\widetilde{\Xi} + \widetilde{\Xi}(M + \overline{P}_\theta(\theta^*))' = -I$.

2.5 Numerical Examples

We use two simulation examples to illustrate the convergence properties. Using (2.5), we take $\varepsilon = 0.01$, $\chi = 1$, $\kappa_1 = -0.271$, $\kappa_2 = 1$, $c_1 = c_2 = 1.5$. For simplicity, we take the additive noise $\mathrm{Pr}(\theta_n, \eta_n) = \mathrm{Pr}(\theta_n) + \eta_n$ and $\mathrm{Pg}(\theta_n, \eta_n) = \mathrm{Pg}(\theta_n) + \eta_n$, where η_n is a sequence of i.i.d. random variables with a standard normal distribution $\mathcal{N}(0, 1)$. In addition, we set the

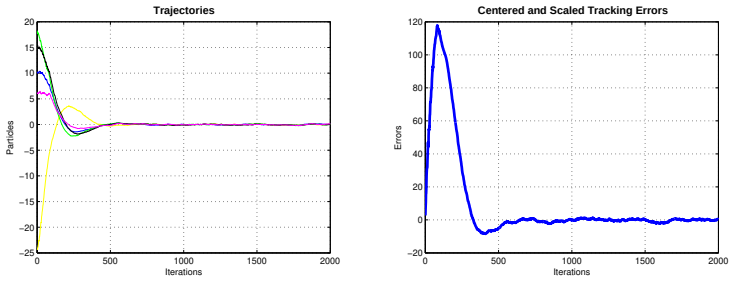


Figure 1: Particle swarm of one-dimensional X using F_1 defined in (2.43).

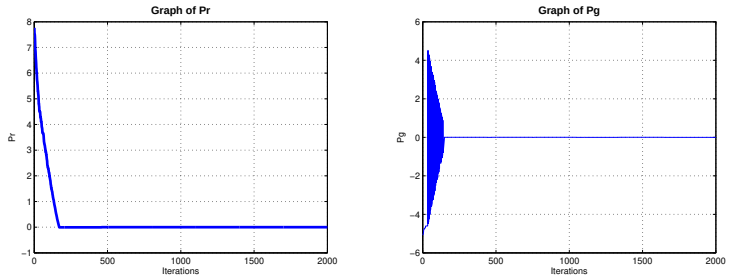


Figure 2: Graphs of Pr and Pg using F_1 defined in (2.43).

number of swarms to be 5.

Example 2.14. *Consider the sphere function:*

$$F_1(x) = \sum_{i=1}^{D} x_i^2, \tag{2.43}$$

where D is the dimension of the variable x. Its global optimum is $(0, 0, \ldots, 0)'$. First, the dimension of X is set to be 1. Figures 1 and 2 show the state trajectories (top) and the centered and scaled errors of the first component θ_n^1 (bottom). The graphs of Pr (top) and Pg (bottom) are also provided.

Next, we consider the 2-dimension case of X. Figures 3 and 4 illustrate the state trajectories (top) and the centered and scaled errors of the first component θ_n^1 (bottom), and the graph of Pr (top) and Pg (bottom), respectively.

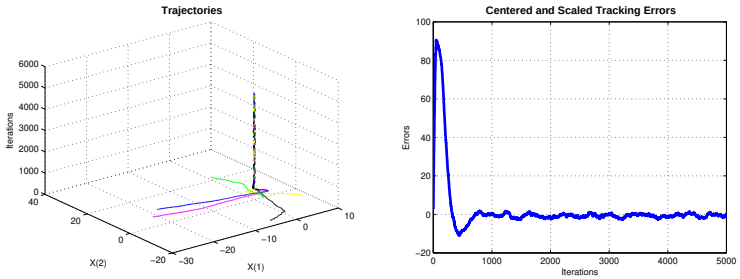


Figure 3: Particle swarm of two-dimensional X using F_1 defined in (2.43).

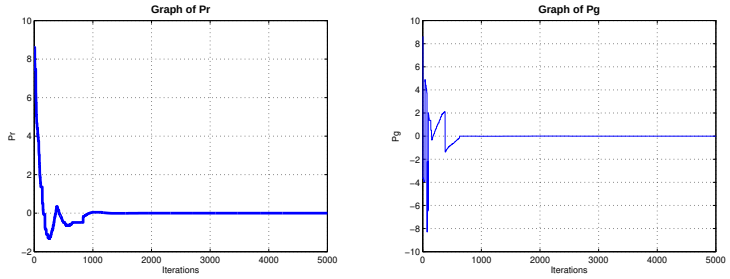


Figure 4: Graphs of Pr and Pg using F_1 defined in (2.43).

Example 2.15. *Consider the Rastrigin function [57]*

$$F_2(x) = 10D + \sum_{i=1}^{D}[x_i^2 - 10\cos(2\pi\ x_i)], \tag{2.44}$$

where D is the dimension of the variable x.

This function has many local minima. Its global optimum is given by $(0, 0, \ldots, 0)'$. Same as Example 2.14, we set the dimension of X to be 1 and 2, respectively. The particle swarm trajectories, the centered and scaled errors of the first component, and graphs of Pr and Pg are given in Figures 5 to 8, respectively.

From these figures, we can conclude that all the swarms converge to a point in the searching space. These results were obtained without assuming that r_1, r_2, Pr, and Pg are fixed. Our numerical results confirm our theoretical findings in Sections 2.3 and 2.4.

Remark 2.16. We use the definition of convergence here that a swarm collapse in which

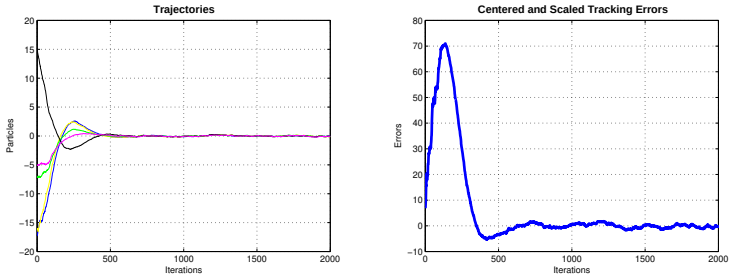


Figure 5: Particle swarm of one-dimensional X using F_2 defined in (2.44).

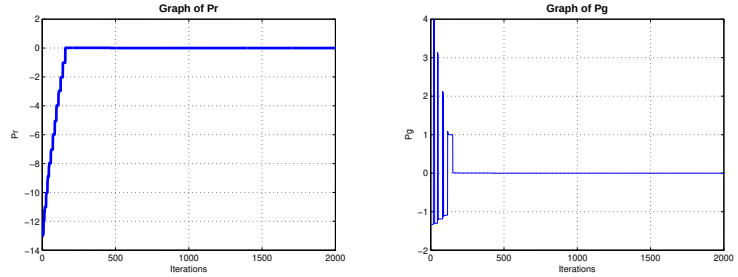


Figure 6: Graphs of Pr and Pg using F_2 defined in (2.44).

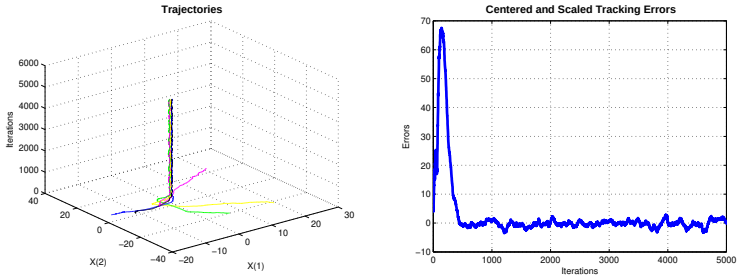


Figure 7: Particle swarm of two-dimensional X using F_2 defined in (2.44).

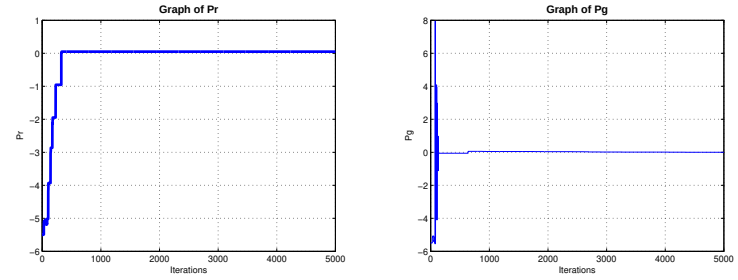


Figure 8: Graphs of Pr and Pg using F_2 defined in (2.44).

all particles have converged to a point in the search space. Sometimes we observe (e.g., in the second example) that the convergence point is not the global or even local optimum. This problem, referred to as premature in literatures, occurs commonly in evolutionary algorithms such as PSOs, genetic algorithms, evolutionary strategies, etc. Based on our numerical experiments, we found that if the cost function is unimodal and with low dimensions, the equilibrium coincides with the proper parameter choice. The problem of under what conditions the equilibrium coincides with the optimum deserves to be carefully studied in the future.

2.6 Further Remarks

In this chapter, we considered a general form of PSO algorithms using a stochastic approximation scheme. Different from the existing results in the literature, we have used weaker assumptions and obtained more general results without depending on empirical work. In addition, we obtained rates of convergence for the PSO algorithms for the first time.

Several research directions may be pursued in the future. We can use stochastic approximation methods to analyze other schemes of PSO, for example, the SPSO2011 considered in [69]. We can set up a stochastic approximation similar to (2.8) and analyze its convergence and convergence rate. Finding ways to systematically choose the parameter values κ_1, κ_2, c_1, and c_2 is a practically challenging problem. One thought is to construct a level two (stochastic) optimization algorithm to select best parameter value in a suitable sense. To proceed in this direction requires careful thoughts and consideration. In addition, we can consider that some parameters such as χ, κ_1, etc. are not fixed but change randomly during iterations or change owing to some random environment change (for example, see [134]). The problem to study is to analyze the convergence and convergence rates in such

a case. Furthermore, using another definition of convergence, i.e., the swarm's best known position P_g approaching (converging to) the optimum of the problem, is another possible study direction.

To conclude, this chapter demonstrated convergence properties of a class of general PSO algorithms and derived the rates of convergence by using a centered and scaled sequence of the iterates. This study opens new arenas for subsequent studies on determining convergence capabilities of different PSO algorithms and parameters.

CHAPTER 3 Infinite Dimensional Regime-Switching SA Algorithms

3.1 Introduction

This chapter is concerned with a class of stochastic approximation (SA) algorithms for tracking the invariant distribution of a Markov chian with has countable state space and is conditioned on another Markov chain also having countable state space. We will evaluate the tracking capability of the SA algorithm in terms of mean squares tracking error, characterize the dynamic behavior of the iterates, reveal the structure of a scaled sequence of tracking errors, and obtain the asymptotic covariance of the associated limit process. Based the discussion in [91], we assume that such a Markov chain with infrequent jumps as a slow Markov chain for simplicity. Since if the parameter changes too fast, there is no chance one can track the time-varying properties using an SA algorithm.

Motivation. This chapter is an extension of the work in [92]. The authors in [92] considered the case that discrete Markov chains have finite state space. We refer the reader to [92] and its references for the background and survey of the problem. In this chapter, we consider the case that the state space of discrete Markov chains is countable. The motivation stems from reduction of computational complexity for large-scale systems, e.g., queueing network models, communication networks, internet traffic controls, and computer systems. Many of these systems are either modeled directly as Markovian systems or can be recast in such a form. In reality, the networks are often quite large with many nodes. Thus the computational complexity is often an important issue and has drawn much attention. We refer the work [93] for more information about discrete-time Markov chains with a countable state space.

Outline. This chapter is devoted to an SA algorithm with constant step size and updates

that are essentially of the form of occupation measures. We focus on the analysis of tracking error bounds. First, we derive mean squares type error bounds using perturbed Lyapunov function methods [90] based on stability analysis. Then we show that an associated system of ODEs with regime switching can be obtained via a combined use of the updated treatment on SA [90] and two-time-scale Markov chains [80, 83]. The system of ODEs with regime switching, different from a single ODE derived from some usual SA algorithms in the existing literatures, is modulated by a continuous-time Markov chain. By this system of switching ODEs, we further analyze the rate of convergence. To do this, we need to examine a sequence of suitably normalized errors. We use a special norm to avoid analysis infinite covariance matrix. We can demonstrate that if the true parameter is a fixed constant, then the norm of this suitable scaled sequence of estimation errors has a Gaussian diffusion limit. Moreover, the limit of the norm is a system of diffusions with regime switching. That means the diffusion coefficient depends on the modulating Markov chain in the limit system,, which reveals the distinctive time-varying nature of the underlying system.

The rest of the chapter is organized as follows. The formulation of the problem is presented in Section 2. Obtaining mean squares error bounds and a weak convergence result of an interpolated sequence of the iterates are showed in Section 3. A norm of suitably scaled tracking error sequence of the iterates and derives a switching diffusion limit are examined in Section 4. Section 5 presents an example of an adaptive discrete stochastic optimization algorithm.

Before proceeding further, a bit of notation is in order. Throughout the chapter, $\mathbb{1}$ denotes an infinite dimensional column vector with all components being 1. For a vector z and a matrix H, we use z' and H' to denote their transposes, and use z^i to denote the ith component of z and H^{ij} to denote the ijth entry of H, respectively. However, for a real

number r, r^k denotes the kth power of r (e.g., ε^k for $\varepsilon > 0$ used in what follows), and $\lfloor r \rfloor$ denotes the integer part of r. K denotes a genetic positive constant whose values may vary for different usage (the conventions $K + K = K$ and $KK = K$ will be used without notice). For a given matrix $H = (h^{ij})_{\infty\times\infty}$ with infinite columns and infinite rows, we define H_a to be a matrix given by $H_a = (\mathbb{1}, H)$. In addition, we use a subscript to index a sequence.

3.2 Formulation

The following conditions are used throughout the chapter. Condition (M) characterizes the time-varying underlying parameter as a Markov chain with infrequent transitions, while condition (S) characterizes the observed signal.

(M) Let $\{\alpha_n\}$ be a discrete-time Markov chain with infinite state space

$$\mathcal{M} = \{\overline{\alpha}_1, \overline{\alpha}_2, \ldots\} \tag{3.1}$$

and transition probability matrix

$$P^\eta = I + \eta\, Q, \tag{3.2}$$

where $\eta > 0$ is a small parameter, I is an infinite dimensional identity matrix, and $Q = (q_{ij})_{\infty\times\infty}$ is a generator of a continuous-time Markov chain (i.e., Q satisfies $q_{ij} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{\infty} q_{ij} = 0$ for each $i = 1, 2, \ldots$). For simplicity, suppose that the initial distribution $P(\alpha_0 = \overline{\alpha}_i) = p_{0,i}$ is independent of η for each $i = 1, 2, \ldots$, where $p_{0,i} \geq 0$ and $\sum_{i=1}^{\infty} p_{0,i} = 1$. Q is irreducible.

(S) Let $\{Y_n\}$ be an infinite state conditional Markov chain (conditioned on the parameter

process). The state space of $\{Y_n\}$ is $\mathcal{S} = \{s_1, s_2, \ldots\}$, where s_i for $i = 1, 2, \ldots$ denotes the ith standard unit vectors, with the ith component being 1 and the rest of the component being 0. For each $\alpha \in \mathcal{M}$, $A(\alpha) = (a_{ij}(\alpha))_{\infty\times\infty}$, the transition probability matrix of Y_n is defined by

$$a_{ij}(\alpha) = P(Y_{n+1} = s_j | Y_n = s_i, \alpha_n = \alpha) = P(Y_1 = s_j | Y_0 = s_i, \alpha_0 = \alpha),$$

where $i, j \in \{1, 2, \ldots\}$. For $\alpha \in \mathcal{M}$, $A(\alpha)$ is irreducible and aperiodic.

By the assumptions, we know that $A(\alpha)$ is irreducible and aperiodic. So there exists a unique stationary distribution $\psi(\alpha) \in \mathbb{R}^{\infty\times 1}$ satisfying

$$\psi'(\alpha) = \psi'(\alpha)A(\alpha) \quad \text{and} \quad \psi'(\alpha)\mathbb{1} = 1.$$

We focus on using an SA algorithm to track the time-varying distribution $\psi(\alpha_n)$ that depends on the underlying Markov chain α_n.

3.2.1 Adaptive Algorithm

We use a stochastic approximation algorithm with constant step size

$$\widehat{\psi}_{n+1} = \widehat{\psi}_n + \varepsilon(Y_{n+1} - \widehat{\psi}_n), \tag{3.3}$$

where ε denotes the step size. This is an adaptive algorithm of least mean squares (LMS) type which can construct a sequence of estimates $\{\widehat{\psi}_n\}$ of the time-varying distribution $\psi(\alpha_n)$,

Define $\widetilde{\psi}_n = \widehat{\psi}_n - \boldsymbol{E}\psi(\alpha_n)$. Then (3.3) can be rewritten as

$$\widetilde{\psi}_{n+1} = \widetilde{\psi}_n - \varepsilon\widetilde{\psi}_n + \varepsilon(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + \boldsymbol{E}(\psi(\alpha_n) - \psi(\alpha_{n+1})). \tag{3.4}$$

Note that $\widehat{\psi}_n$, $\psi(\alpha)$, and hence $\widetilde{\psi}_n$ are infinite column vectors (i.e., they take values in $\mathbb{R}^{\infty\times 1}$).

The underlying parameter α_n is called a *hypermodel* in [91]. Although the dynamics of the hypermodel α_n is used in our analysis, it does not explicitly enter the implementation of the LMS algorithm (3.3).

Now we will derive a mean square error bound by examining an interpolated sequence of the iterations, and derive a limit result for a scaled sequence in the following sections.

3.3 Asymptotic Properties

3.3.1 Mean Square Error

We consider a mean square estimate for $\boldsymbol{E}|\widetilde{\psi}_n|^2 = \boldsymbol{E}|\widehat{\psi}_n - \boldsymbol{E}\psi(\alpha_n)|^2$ first. Lyapunnov-type functions are often required to analyze SA algorithms for proving stability, see [73, 90]. In what follows, we establish the desired estimate via a stability argument using the perturbed Lyapunov function method [90]. Use $\boldsymbol{E}_n$ to denote the conditional expectation with respect to $\mathcal{F}_n$, the σ-algebra generated by $\{Y_k, \alpha_k : k \leq\ n\}$.

Theorem 3.1. *Assume* (M) *and* (S). *In addition, suppose that* $\eta^2 \ll \varepsilon$. *Then for sufficiently large* n,

$$\boldsymbol{E}|\widetilde{\psi}_n|^2 = O\left(\varepsilon + \eta + \frac{\eta^2}{\varepsilon}\right). \tag{3.5}$$

Proof. Define $V(x) = (x'x)/2$. Direct calculations lead to

$$\begin{aligned} \boldsymbol{E}_n V(\widetilde{\psi}_{n+1}) - V(\widetilde{\psi}_n) = \boldsymbol{E}_n\{\widetilde{\psi}_n'[-\varepsilon\widetilde{\psi}_n + \varepsilon(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + \boldsymbol{E}[\psi(\alpha_n) - \psi(\alpha_{n+1})]]\} \\ + \boldsymbol{E}_n|-\varepsilon\widetilde{\psi}_n + \varepsilon(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + \boldsymbol{E}[\psi(\alpha_n) - \psi(\alpha_{n+1})]|^2. \end{aligned} \tag{3.6}$$

By the Markovian assumption and the structure of the transition probability matrix given by (3.2),

$$\begin{aligned} \boldsymbol{E}_n[\psi(\alpha_n) - \psi(\alpha_{n+1})] &= \boldsymbol{E}[\psi(\alpha_n) - \psi(\alpha_{n+1})|\alpha_n] \\ &= \sum_{i=1}^{\infty} \boldsymbol{E}[\psi(\overline{\alpha}_i) - \psi(\alpha_{n+1})|\alpha_n = \overline{\alpha}_i] I_{\{\alpha_n = \overline{\alpha}_i\}} \\ &= \sum_{i=1}^{\infty} \left[\psi(\overline{\alpha}_i) - \sum_{j=1}^{\infty} \psi(\overline{\alpha}_j) p_{ij}^{\eta}\right] I_{\{\alpha_n = \overline{\alpha}_i\}} \\ &= -\eta \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \psi(\overline{\alpha}_j) q_{ij} I_{\{\alpha_n = \overline{\alpha}_i\}} \\ &= O(\eta), \end{aligned} \tag{3.7}$$

moreover, detailed computation also shows that

$$\boldsymbol{E}_n|\psi(\alpha_n) - \psi(\alpha_{n+1})|^2 = O(\eta). \tag{3.8}$$

Owing to (3.2), the transition probability matrix P^{η} is independent of time n. As a result, the k-step transition probability depends only on the time lags and can be denoted by $(P^{\eta})^k$. By an elementary inequality, we have $|\widetilde{\psi}_n| = |\widetilde{\psi}_n| \cdot 1 \leq (|\widetilde{\psi}_n|^2 + 1)/2$. Thus,

$$O(\eta)|\widetilde{\psi}_n| \leq O(\eta)(V(\widetilde{\psi}_n) + 1).$$

Noting that the sequence of signals $\{Y_n\}$ is bounded, the boundedness of $\{\widehat{\psi}_n\}$, and $O(\eta\varepsilon) = O(\eta^2 + \varepsilon^2)$ via the elementary inequality $ab \leq (a^2 + b^2)/2$ for any real numbers a

and b, the estimate (4.14) yields

$$\begin{aligned}
&\boldsymbol{E}_n|-\varepsilon\widetilde{\psi}_n+\varepsilon(Y_{n+1}-\boldsymbol{E}\psi(\alpha_n))+\boldsymbol{E}[\psi(\alpha_n)-\psi(\alpha_{n+1})]|^2\\
&\le K\boldsymbol{E}_n\Big[\varepsilon^2|\widetilde{\psi}_n|^2+\varepsilon^2|Y_{n+1}-\boldsymbol{E}\psi(\alpha_n)|^2+\varepsilon^2|\widetilde{\psi}'_n\boldsymbol{E}(Y_{n+1}-\boldsymbol{E}\psi(\alpha_n))|\\
&\qquad\qquad +\varepsilon|\widetilde{\psi}'_n\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1}))|+\varepsilon|(Y_{n+1}-\boldsymbol{E}\psi(\alpha_n))'E(\psi(\alpha_n)-\psi(\alpha_{n+1}))|\Big]\\
&\quad+|\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1}))|^2\\
&=O(\varepsilon^2+\eta^2)(V(\widetilde{\psi}_n)+1)+|\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1}))|^2
\end{aligned}\tag{3.9}$$

and

$$\begin{aligned}
&\boldsymbol{E}_n\widetilde{\psi}'_n[-\varepsilon\widetilde{\psi}_n+\varepsilon(Y_{n+1}-\boldsymbol{E}\psi(\alpha_n))+\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1}))]\\
&\quad=-2\varepsilon V(\widetilde{\psi}_n)+\varepsilon\boldsymbol{E}_n\widetilde{\psi}'_n(Y_{n+1}-\boldsymbol{E}\psi(\alpha_n))+\boldsymbol{E}_n\widetilde{\psi}'_n\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1})).
\end{aligned}\tag{3.10}$$

Using (3.9) and (3.10) in (4.12) together with (4.13), we obtain

$$\begin{aligned}
&\boldsymbol{E}_nV(\widetilde{\psi}_{n+1})-V(\widetilde{\psi}_n)\\
&\quad=-2\varepsilon V(\widetilde{\psi}_n)+\varepsilon\boldsymbol{E}_n\widetilde{\psi}'_n(Y_{n+1}-E\psi(\alpha_n))+\boldsymbol{E}_n\widetilde{\psi}'_n\boldsymbol{E}(\psi(\alpha_n)-\psi(\alpha_{n+1}))\\
&\qquad+O(\varepsilon^2+\eta^2)(V(\widetilde{\psi}_n)+1).
\end{aligned}\tag{3.11}$$

To obtain the desired estimate, we need to "average out" the second to the fourth terms on the right-hand side of (3.11). To do so, for any $0<T<\infty$, we define the following perturbations:

$$\begin{aligned}
V_1^\eta(\widetilde{\psi},n)&=\varepsilon\sum_{j=n}^{T/\eta}\widetilde{\psi}'\boldsymbol{E}_n(Y_{j+1}-\boldsymbol{E}\psi(\alpha_j)),\\
V_2^\eta(\widetilde{\psi},n)&=\sum_{j=n}^{T/\eta}\widetilde{\psi}'\boldsymbol{E}(\psi(\alpha_j)-\psi(\alpha_{j+1})).
\end{aligned}\tag{3.12}$$

In the above and hereafter, T/η is understood to be $\lfloor T/\eta\rfloor$, i.e., the integer part of T/η.

Throughout the rest of the proof, we often need to use the notion of fixed-α processes. For example, by $Y_j(\alpha)$ for $n\le j\le O(1/\eta)$, we mean a process in which $\alpha_j=\alpha$ is fixed for

all j with $n \leq j \leq O(1/\eta)$.

For $V_1^\eta(\widetilde{\psi}, n)$ defined in (3.12),

$$\begin{aligned}\left|\sum_{j=n}^{T/\eta} \boldsymbol{E}_n[Y_{j+1} - \psi(\alpha_j)]\right| \leq & \left|\sum_{j=n}^{T/\eta} \boldsymbol{E}_n[Y_{j+1} - \boldsymbol{E}Y_{j+1}]\right| \\ & + \left|\sum_{j=n}^{T/\eta} [\boldsymbol{E}Y_{j+1} - \boldsymbol{E}\psi(\alpha_j)]\right|.\end{aligned} \tag{3.13}$$

Using the ϕ-mixing property of $\{Y_j\}$ (see [76, p. 166]),

$$\left|\sum_{j=n}^{T/\eta} \boldsymbol{E}_n[Y_{j+1} - \boldsymbol{E}Y_{j+1}]\right| \leq K \leq \infty \qquad \text{uniformly in } n. \tag{3.14}$$

We can also show

$$\left|\sum_{j=n}^{T/\eta} [\boldsymbol{E}Y_{j+1} - \boldsymbol{E}\psi(\alpha_j)]\right| < \infty. \tag{3.15}$$

Thus, using (3.13)-(4.21), for each $\widetilde{\psi}$,

$$|V_1^\eta(\widetilde{\psi}, n)| \leq O(\varepsilon)(V(\widetilde{\psi}) + 1). \tag{3.16}$$

By virtue of the definition of $V_2^\eta(\cdot)$ and (3.2), it follows that there exists an N_η for all $n \geq N_\eta$ such that

$$\begin{aligned}|V_2^\eta(\widetilde{\psi}, n)| &= \left|\sum_{j=n}^{T/\eta} \widetilde{\psi}'[\boldsymbol{E}(\psi(\alpha_j) - \psi(\alpha_{j+1}))]\right| \\ &= |\widetilde{\psi}'\boldsymbol{E}[\psi(\alpha_n) - \psi(\alpha_{T/\eta})]| \\ &\leq |\widetilde{\psi}|O(\eta) \\ &\leq O(\eta)(V(\widetilde{\psi}) + 1).\end{aligned} \tag{3.17}$$

We next show that they result in the desired cancellation in the error estimate. Note that

$$\begin{aligned}&\boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_{n+1}, n+1) - V_1^\eta(\widetilde{\psi}_n, n)\\ &\quad = \boldsymbol{E}_n[V_1^\eta(\widetilde{\psi}_{n+1}, n+1) - V_1^\eta(\widetilde{\psi}_n, n+1)] + \boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_n, n+1) - V_1^\eta(\widetilde{\psi}_n, n).\end{aligned} \tag{3.18}$$

It can be seen that

$$\boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_n, n+1) - V_1^\eta(\widetilde{\psi}_n, n) = -\varepsilon \boldsymbol{E}_n \widetilde{\psi}_n'(Y_{n+1} - E\psi(\alpha_n)) \tag{3.19}$$

and

$$\begin{aligned}&\boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_{n+1}, n+1) - \boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_n, n+1)\\ &= \varepsilon \sum_{j=n+1}^{T/\eta} \boldsymbol{E}_n \widetilde{\psi}_{n+1}' \boldsymbol{E}_{n+1}(Y_{j+1} - \boldsymbol{E}\psi(\alpha_j)) - \varepsilon \sum_{j=n+1}^{T/\eta} \boldsymbol{E}_n \widetilde{\psi}_n' \boldsymbol{E}_{n+1}(Y_{j+1} - \boldsymbol{E}\psi(\alpha_j))\\ &= \varepsilon \sum_{j=n+1}^{T/\eta} \boldsymbol{E}_n (\widetilde{\psi}_{n+1} - \widetilde{\psi}_n)' \boldsymbol{E}_{n+1}(Y_{j+1} - \boldsymbol{E}\psi(\alpha_j))\\ &= \varepsilon \sum_{j=n+1}^{T/\eta} \boldsymbol{E}_n [-\widetilde{\psi}_n + \varepsilon(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + E(\psi(\alpha_n) - \psi(\alpha_{n+1}))]' \boldsymbol{E}_{n+1}[Y_{j+1} - E\psi(\alpha_j)]\\ &= O(\varepsilon^2)(V(\widetilde{\psi}_n) + 1) + O(\varepsilon\eta) = O(\varepsilon^2)(V(\widetilde{\psi}_n) + 1) + O(\eta^2).\end{aligned} \tag{3.20}$$

In the above, we have used $O(\varepsilon\eta) = O(\varepsilon^2 + \eta^2)$, (3.4), and (4.12) to obtain

$$\begin{aligned}|\boldsymbol{E}_n[\widetilde{\psi}_{n+1} - \widetilde{\psi}_n]| &\le \varepsilon \boldsymbol{E}_n |\widetilde{\psi}_n| + \varepsilon \boldsymbol{E}_n |Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)| + O(\eta)\\ &= O(\varepsilon)(V(\widetilde{\psi}_n) + 1) + O(\eta).\end{aligned} \tag{3.21}$$

Thus

$$\begin{aligned}&\boldsymbol{E}_n V_1^\eta(\widetilde{\psi}_{n+1}, n+1) - V_1^\eta(\widetilde{\psi}_n, n)\\ &\quad = -\boldsymbol{E}_n \widetilde{\psi}_n'(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + O(\varepsilon^2)(V(\widetilde{\psi}_n) + 1) + O(\eta^2).\end{aligned} \tag{3.22}$$

Analogous estimates yield that

$$\begin{aligned}
&\boldsymbol{E}_n V_2^\eta(\widetilde{\psi}_{n+1}, n+1) - \boldsymbol{E}_n V_2^\eta(\widetilde{\psi}_n, n+1) \\
&\quad = \sum_{j=n+1}^{T/\eta} \boldsymbol{E}_n(\widetilde{\psi}_{n+1} - \widetilde{\psi}_n)' \boldsymbol{E}(\psi(\alpha_j) - \psi(\alpha_{j+1})) \\
&\quad = O(\varepsilon\eta)(V(\widetilde{\psi}_n) + 1) + O(\eta^2) = O(\eta^2 + \varepsilon^2)(V(\widetilde{\psi}_n) + 1),
\end{aligned} \tag{3.23}$$

and that

$$\boldsymbol{E}_n V_2^\eta(\widetilde{\psi}_n, n+1) - V_2^\eta(\widetilde{\psi}_n, n) = -\widetilde{\psi}_n' \boldsymbol{E}(\psi(\alpha_n) - \psi(\alpha_{n+1})). \tag{3.24}$$

Thus,

$$\begin{aligned}
&\boldsymbol{E}_n V_2^\eta(\widetilde{\psi}_{n+1}, n+1) - V_2^\eta(\widetilde{\psi}_n, n) \\
&\quad = -\widetilde{\psi}_n' \boldsymbol{E}(\psi(\alpha_n) - \psi(\alpha_{n+1})) + O(\varepsilon^2 + \eta^2)(V(\widetilde{\psi}_n) + 1).
\end{aligned} \tag{3.25}$$

Redefine V_1^η and V_2^η with T/η replaced by ∞. Estimates (3.13)-(3.25) still hold.

Define

$$W(\widetilde{\psi}, n) = V(\widetilde{\psi}) + V_1^\eta(\widetilde{\psi}, n) + V_2^\eta(\widetilde{\psi}, n).$$

Then, using the above estimates, we have

$$\begin{aligned}
&\boldsymbol{E}_n W(\widetilde{\psi}_{n+1}, n+1) - W(\widetilde{\psi}_n, n) \\
&\quad = \boldsymbol{E}_n V(\widetilde{\psi}_{n+1}) - V(\widetilde{\psi}_n) + \boldsymbol{E}_n[V_1^\eta(\widetilde{\psi}_{n+1}, n+1) - V_1^\eta(\widetilde{\psi}_n, n)] \\
&\qquad + \boldsymbol{E}_n[V_2^\eta(\widetilde{\psi}_{n+1}, n+1) - V_2^\eta(\widetilde{\psi}_n, n)] \\
&\quad = -2\varepsilon V(\widetilde{\psi}_n) + O(\varepsilon^2 + \eta^2)(V(\widetilde{\psi}_n) + 1).
\end{aligned} \tag{3.26}$$

This, together with (4.22) and (3.17) and T/η replaced by ∞, implies

$$\begin{aligned}
&\boldsymbol{E}_n W(\widetilde{\psi}_{n+1}, n+1) - W(\widetilde{\psi}_n, n) \\
&\qquad \le -2\varepsilon W(\widetilde{\psi}_n, n) + O(\varepsilon^2 + \eta^2)(W(\widetilde{\psi}_n, n) + 1).
\end{aligned} \tag{3.27}$$

Choose ε and η small enough so that there is a $\lambda > 0$ satisfying

$$-2\varepsilon + O(\eta^2) + O(\varepsilon^2) \leq -\lambda\varepsilon.$$

Then, we get

$$\boldsymbol{E}_n W(\widetilde{\psi}_{n+1}, n+1) \leq (1-\lambda\varepsilon) W(\widetilde{\psi}_n, n) + O(\varepsilon^2 + \eta^2). \tag{3.28}$$

Taking the expectation and iterating on the resulting inequality yields

$$\begin{aligned} \boldsymbol{E}W(\widetilde{\psi}_{n+1}, n+1) &\leq (1-\lambda\varepsilon)^{n-N_\eta} \boldsymbol{E}W(\widetilde{\psi}_0, 0) + \sum_{j=N_\eta}^{n} (1-\lambda\varepsilon)^{j-N_\eta} O(\varepsilon^2+\eta^2) \\ &\leq (1-\lambda\varepsilon)^{n-N_\eta} \boldsymbol{E}W(\widetilde{\psi}_0, 0) + O\left(\varepsilon + \frac{\eta^2}{\varepsilon}\right) \end{aligned} \tag{3.29}$$

By taking n large enough, we can make $(1-\lambda\varepsilon)^{n-N_\eta} = (\varepsilon)$. Then

$$\boldsymbol{E}W(\widetilde{\psi}_{n+1}, n+1) \leq O\left(\varepsilon + \frac{\eta^2}{\varepsilon}\right). \tag{3.30}$$

Finally, applying (4.22) and (3.17) again, replacing $W(\widetilde{\psi}, n)$ by $V(\widetilde{\psi})$ adds another $O(\eta)$ term. Thus we obtain

$$\boldsymbol{E}V(\widetilde{\psi}_{n+1}) \leq O\left(\varepsilon + \eta + \frac{\eta^2}{\varepsilon}\right). \tag{3.31}$$

This concludes the proof. □

Since our adaptive algorithm can track the time-varying parameter, the ratio η/ε must not be large. Given the order-of-magnitude estimate $O(\varepsilon + \eta + \eta^2/\varepsilon)$, to balance the two terms ε and η^2/ε, we need to choose $\eta = O(\varepsilon)$. By Theorem 3.1, we obtain the following result.

Corollary 3.2. *Under the conditions of Theorem* 3.1*, if* $\eta = O(\varepsilon)$*, then for sufficiently*

large n, $\boldsymbol{E}|\widetilde{\psi}_n|^2 = O(\varepsilon)$.

3.3.2 Limit System of Regime-Switching ODEs

Next, we try to derive a limit system for an interpolated sequence of the iterates. We consider the case $\eta = O(\varepsilon)$. For notational simplicity, we use $\eta = \varepsilon$. For $0 < T < \infty$, we construct a sequence of piecewise constant interpolation of the stochastic approximation iterates $\widehat{\psi}_n$ as

$$\widehat{\psi}^{\varepsilon}(t) = \widehat{\psi}_n, \quad t \in [\varepsilon n, \varepsilon n + \varepsilon). \tag{3.32}$$

The process $\widehat{\psi}^{\varepsilon}(\cdot)$ so defined is in $D([0,T];\mathbb{R}^{\infty})$, which is the space of functions defined on $[0,T]$ taking values in $\mathbb{R}^{\infty}$ that are right continuous, have left limits, and are endowed with the Skorohod topology. We implement the analysis using weak convergence methods. The application of weak convergence ideas usually requires proof of tightness and the characterization of the limit processes, which is a system of ODEs modulated by a continuous-time Markov chain.

Lemma 3.3. *Under conditions* (M) *and* (S), $\{\psi^{\varepsilon}(\cdot)\}$ *is tight in* $D([0,T];\mathbb{R}^{\infty})$.

Proof. By using the tightness criteria [77, p. 47], it suffices to verify that for any $\delta > 0$ and $0 < s \le \delta$,

$$\lim_{\delta\to 0}\limsup_{\varepsilon\to 0} \boldsymbol{E}[\sup_{0\le s\le\delta} \boldsymbol{E}_t^{\varepsilon}|\widehat{\psi}^{\varepsilon}(t+s) - \widehat{\psi}^{\varepsilon}(t)|^2] = 0. \tag{3.33}$$

Note that

$$\begin{aligned}\widehat{\psi}^{\varepsilon}(t+s) - \widehat{\psi}^{\varepsilon}(t) &= \widehat{\psi}_{(t+s)/\varepsilon} - \widehat{\psi}_{t/\varepsilon}\\ &= \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} (Y_{k+1} - \widehat{\psi}_k).\end{aligned} \tag{3.34}$$

Note also that both the iterates and the observations are bounded uniformly. Then the

boundedness of $\{Y_k\}$ and $\{\widehat{\psi}_k\}$ implies that

$$\begin{aligned}
&\boldsymbol{E}_t^\varepsilon|\widehat{\psi}^\varepsilon(t+s)-\widehat{\psi}^\varepsilon(t)|^2\\
&=\boldsymbol{E}_t^\varepsilon\left[\varepsilon\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}(Y_{k+1}-\widehat{\psi}_k)'\right]\left[\varepsilon\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}(Y_{j+1}-\widehat{\psi}_j)'\right]\\
&=\varepsilon^2\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1}\sum_{j=t/\varepsilon}^{(t+s)/\varepsilon-1}\boldsymbol{E}(Y_{k+1}-\widehat{\psi}_k)'(Y_{j+1}-\widehat{\psi}_j)\\
&\le\ K\varepsilon^2\left(\frac{t+s}{\varepsilon}-\frac{t}{\varepsilon}\right)\\
&=K((t+s)-t)^2=O(\delta^2).
\end{aligned}\tag{3.35}$$

Then (3.33) follows, so the desired tightness follows. □

3.4 Limit of Modulating Markov Chain

Consider the Markov chain $\{\alpha_n\}$. Regarding the probability vector and the n-step transition probability matrix, we have the following approximation results.

Suppose that α_n^μ is a discrete Markov chain which has l ($l<\infty$) interconnected subspaces such that its state spaces is given by

$$\mathcal{M}=\mathcal{M}_1\cup\mathcal{M}_2\cup\ldots\mathcal{M}_l,\tag{3.36}$$

where $\mathcal{M}_i=\{s_{i1},s_{i2},\ldots\}$ for $i=1,\ldots,l$. Within each subset the transitions take place an order of magnitude more frequent than that of among different subsets. The transition probability matrix is

$$P^\mu=P+\mu Q,\tag{3.37}$$

where

$$P=\mathrm{diag}(P^1,\ldots,P^l)\tag{3.38}$$

is the usual notation of diagonal block matrix with entries of appropriate dimensions.

The asymptotic expansions of the probability vector is constructed as

$$\begin{aligned} p_k^{\mu} &= (p_k^{\mu,11}, \ldots, p_k^{\mu,21}, \ldots, \ldots, p_k^{\mu,l1}, \ldots) \\ &= (P(\alpha_k^{\mu} = s_{11}), \ldots, P(\alpha_k^{\mu} = s_{21}), \ldots, \ldots, P(\alpha_k^{\mu} = s_{l1}), \ldots). \end{aligned} \tag{3.39}$$

We often partition an infinite-dimensional vector φ in accordance with the decomposition of the state space given by (3.36) as

$$\varphi = (\varphi^1, \ldots, \varphi^l) \quad \text{where } \varphi^i = (\varphi^{i1}, \varphi^{i2}, \ldots). \tag{3.40}$$

That is, φ^i is an infinite-dimensional vector corresponding to the subspace $\mathcal{M}_i$.

Following the approach in [93], consider the spaces

$$\ell_1 = \{(\varphi^1, \ldots, \varphi^l) : 1 \le i \le l, \varphi^{ik} \in \mathbb{R} \quad \text{for each} \quad k \in \mathbb{N}, \quad \text{and} \quad \sum_{i=1}^{l} \sum_{k=1}^{\infty} |\varphi^{ik}| < \infty\},$$

$$\ell_\infty = \{(\varphi^1, \ldots, \varphi^l) : 1 \le i \le l, \varphi^{ik} \in \mathbb{R} \quad \text{for each} \quad k \in \mathbb{N}, \quad \text{and} \quad \sup_{1 \le i \le l} \sup_{1 \le k \le \infty} |\varphi^{ik}| < \infty\},$$

equipped with the norms

$$\|\varphi\|_1 = \sum_{i=1}^{l} \sum_{k=1}^{\infty} |\varphi^{ik}|, \quad \text{and} \quad \|\varphi\|_\infty = \sup_{1 \le i \le l} \sup_{1 \le k < \infty} |\varphi^{ik}|,$$

respectively; see Huston and Pym [81, p. 11]. For a linear operator A defined on these spaces, we use its induced norm $\|A\| = \sup_{\|x\|=1} \|Ax\|$, where $\|\cdot\|$ is either the norm $\|\cdot\|_1$ or $\|\cdot\|_\infty$. It is easily seen that $p_k^{\mu} \in \ell_1$ and that for each $i (1 \le i \le l)$ and each $k (1 \le k < \infty)$,

$p_k^{\mu,ij} \geq 0$, and $\sum_{i=1}^{l} \sum_{j=1}^{\infty} p_k^{\mu,ij} = 1$. It is also well known that p_k^{μ} satisfies the equation

$$p_{k+1}^{\mu} = p_k^{\mu} P^{\mu}. \tag{3.41}$$

Assume that the initial probability vector p_0 is independent of μ,

$$p_0^{\mu} = p_0 = (p_0^{11}, \ldots, p_0^{21}, \ldots, p_0^{l1}, \ldots)$$

such that

$$p_0^{ij} \geq 0, \quad \text{and} \quad \sum_{i=1}^{l} \sum_{j=1}^{\infty} p_0^{ij} = 1. \tag{3.42}$$

In addition,

$$\sup_{0 \leq k \leq T/\mu} \|p_k^{\mu}\|_{\infty} \leq 1 \text{ and } \sup_{0 \leq k \leq T/\mu} \|p_k^{\mu}\|_1 = \sup_{0 \leq k \leq T/\mu} \sum_{i=1}^{l} \sum_{j=1}^{\infty} p_k^{ij} = 1,$$

since it is a probability vector.

We will use the following two assumptions.

(A1) Let P^{μ} be given by (3.37) with P specified in (3.38), P^{μ} and P are transition probability matrices; for each $i \leq l$, P^i is irreducible and aperiodic.

(A2) For each $i = 1, \ldots, l$, there is a $0 \leq \lambda_i < 1$ such that for $k \geq 1$,

$$\|(P^i)^k - \mathbb{1} v^i\|_{\infty} \leq K(\lambda_i)^k, \tag{3.43}$$

where $v^i = (v^{i1}, v^{i2}, \ldots)$ is the stationary distribution corresponding to the transition matrix P^i.

We have the following lemma.

Lemma 3.4. *Under the conditions* (A1) *and* (A2)*,*

$$\sup_{0\leq k\leq T/\mu}\left\|p_k^\mu-\left[\sum_{j=0}^n\mu^j u_j(\mu\ k)+\sum_{j=0}^n\mu^j v_j(k)\right]\right\|_\infty=O(\mu^{n+1}).$$

where

$$u_0(t)(P-I)=0,\quad u_\ell(t)(P-I)=\sum_{\imath=1}^{\ell}\tfrac{1}{\imath!}\tfrac{d^\imath u_{\ell-\imath}(t)}{dt^\imath}-u_{\ell-1}(t)Q,\ell=1,\ldots,n,$$
$$v_0(k+1)=v_0(k)P,\quad v_\ell(k+1)=v_\ell(k)P+v_{\ell-1}(k)Q,\ell=1,\ldots,n,$$

and

$$u_0(0)+v_0(0)=p_0,\quad u_\ell(0)=-v_\ell(0),\ell=1,\ldots,n.$$

In addition,

$$P_k^\mu=\sum_{\ell=0}^n\mu^\ell\ U_\ell(\mu k)+\sum_{\ell=0}^n\mu^\ell V_\ell(k)+O(\mu^{n+1}),\tag{3.44}$$

uniformly in $k=1,\ldots,T/\mu$*, where*

$$U_0(t)(P-I)=0,\quad U_\ell(t)(P-I)=\sum_{\imath=1}^{\ell}\frac{1}{\imath\ !}\frac{dU_{\ell-\imath}(t)}{dt^\imath}-U_{\ell-1}(t)G,\ell=1,\ldots,n,$$

$$V_0(k+1)=V_0(k)P,\quad V_\ell(k+1)=V_\ell(k)P+V_{\ell-1}(k)G,\ell=1,\ldots,n,$$

and

$$U_0(0)+V_0(0)=I,\quad U_\ell(0)=-V_\ell(0),\ell=1,\ldots,n.$$

Proof. See the proofs of Theorems 2.6 and 2.7 in [93]. □

Lemma 3.5. *Under conditions (A1) and (A2),* $\overline{\alpha}^\mu(\cdot)$ *converges weakly to* $\overline{\alpha}(\cdot)$*, a Markov*

chain generated by $\overline{Q}$ *defined by*

$$\overline{Q} = vQ\widetilde{\mathbb{1}} = \operatorname{diag}(v^1, \ldots, v^l)\, Q \operatorname{diag}(\mathbb{1}, \mathbb{1}, \ldots, \mathbb{1}). \tag{3.45}$$

Proof. See the proof of Theorem 2.10 in [93]. □

With the above two lemmas, we can now derive a result that will be used in the subsequence analysis. The proof is essentially an application of the above lemmas.

Proposition 3.6. *Assume (M). Choose* $\eta = \varepsilon$ *and consider the Markov chain* α_n. *Then the following assertions hold*

- *Denote* $p_n^\varepsilon = (P(\alpha_n = \overline{\alpha}_1), \ldots, P(\alpha_n = \overline{\alpha}_2), \ldots)$. *Then*

$$\begin{aligned} &p_n^\varepsilon = z(t) + O(\varepsilon), \quad z(t) \in \mathbb{R}^{1\times\infty}, \\ &\frac{dz(t)}{dt} = z(t)Q, \quad z(0) = p_0, \\ &(P^\varepsilon)^n = Z(t) + O(\varepsilon), \\ &\frac{dZ(t)}{dt} = Z(t)Q, \quad Z(t) = I. \end{aligned} \tag{3.46}$$

- *Define the continuous-time interpolation of* α_n^ε *by* $\alpha^\varepsilon(t) = \alpha_n$ *if* $t \in [n\varepsilon, n\varepsilon + \varepsilon)$. *Then* $\alpha^\varepsilon(\cdot)$ *converges weakly to* $\alpha(\cdot)$, *which is a continuous-time Markov chain generated by* Q.

Proof. Observe that the identity matrix in (3.2) can be written as

$$I = \operatorname{diag}(1, 1, \ldots) \in \mathbb{R}^{\infty\times\infty}.$$

Each of the 1's can be thought as a 1×1 "transition matrix". Note that under the con-

ditions for the Markov chain α_n, the $\text{diag}(v^1, \ldots, v^l)$ defined in (3.45) becomes $I \in \mathbb{R}^{\infty\times\infty}$, and $\text{diag}(\mathbb{1}, \mathbb{1}, \ldots)$ in (3.45) is also I. Moreover, the $\overline{Q}$ defined in (3.45) is now simply Q. Straightforward applications of Lemma 3.4 and Lemma 3.5 then yield the desired results. □

3.4.1 Limit Differential Equations

Consider the pair $(\widehat{\psi}^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$. By Proposition 3.6 and Lemma 3.3 together with the Cramér-Wold device [76, p. 48], we know $\{(\widehat{\psi}^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))\}$ is tight in $D([0,T]; \mathbb{R}^\infty \times \mathcal{M})$ for $T > 0$. By virtue of Prohorov's theorem, we can extract convergent subsequences. For notational simplicity, we still index the subsequence by ε and denote the limit by $\widehat{\psi}(\cdot)$. By virtue of the Skorohod representation, $\widehat{\psi}^\varepsilon(\cdot)$ converges to $\widehat{\psi}(\cdot)$ w.p.1, and the convergence is uniform on any compact interval. We proceed to characterize the limit $\widehat{\psi}(\cdot)$. The result is stated in the following theorem.

Theorem 3.7. *Under conditions* (M) *and* (S), $(\widehat{\psi}^\varepsilon(\cdot), \alpha^\varepsilon(\cdot))$ *converges weakly to* $(\widehat{\psi}(\cdot), \alpha(\cdot))$, *which is a solution of the following switching ODE:*

$$\frac{d}{dt}\widehat{\psi}(t) = \psi(\alpha(t)) - \widehat{\psi}(t), \quad \widehat{\psi}(0) = \widehat{\psi}_0. \tag{3.47}$$

Proof. To obtain the desired limit, we prove that the limit $(\widehat{\psi}(\cdot), \alpha(\cdot))$ is the solution of the martingale problem with operator L_1 given by

$$L_1 f(x, \overline{\alpha}_i) = \nabla\ f'(x, \overline{\alpha}_i)(\psi(\overline{\alpha}_i) - x) + Qf(x, \cdot)(\overline{\alpha}_i) \quad \text{for each} \quad \overline{\alpha}_i \in \mathcal{M}, \tag{3.48}$$

where

$$Q\ f(x,\cdot)(\overline{\alpha}_i) = \sum_{j\in\mathcal{M}} q_{ij} f(x,\overline{\alpha}_j) = \sum_{i\neq j} q_{ij}[f(x,\overline{\alpha}_j) - f(x,\overline{\alpha}_i)] \quad \text{for each} \quad \overline{\alpha}_i \in \mathcal{M},$$

and for each $\overline{\alpha}_i \in \mathcal{M}$, $f(\cdot,\overline{\alpha}_i) \in C_0^2$, the class of functions that are twice continuously differentiable with compact support. In the above, $\nabla\ f(x,\overline{\alpha}_i)$ denotes the gradient of $f(x,\overline{\alpha}_i)$ with respect to x. Using an argument as in [80, Lemma 7.18], it can be shown that the martingale problem associated with the operator L_1 has a unique solution. To show that the limit $(\widehat{\psi}(\cdot),\alpha(\cdot))$ is a solution of the martingale problem with operator L_1, it suffices to show that for each $\overline{\alpha}_i \in \mathcal{M}$ and any $f(\cdot,\overline{\alpha}_i) \in C_0^2$, $f(x(t)-\alpha(t)) - f(x(0),\alpha(0)) - \int_0^t L_1 f(x(s),\alpha(s))ds$ is a martingale. To verify this, we need only to show that for any positive integer ℓ_0, any $t>0$, $s>0$, and $0<t_j\leq t$, and any bounded and continuous function $h_j(\cdot,\overline{\alpha}_i)$ for each $\overline{\alpha}_i\in\mathcal{M}$ with $j\leq\ell_0$,

$$\begin{aligned} &\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}(t_j),\alpha(t_j)) \\ &\quad\times\left[f(\widehat{\psi}(t+s),\alpha(t+s)) - f(\widehat{\psi}(t),\alpha(t)) - \int_t^{t+s} L_1 f(\widehat{\psi}(u),\alpha(u))du\right] = 0. \end{aligned} \tag{3.49}$$

To verify (3.49), we work with the processes indexed by ε and prove that the above equation holds as $\varepsilon\to\ 0$.

First by the weak convergence of $(\widehat{\psi}^\varepsilon(\cdot),\alpha^\varepsilon(\cdot))$ to $(\widehat{\psi}(\cdot),\alpha(\cdot))$ and the Skorohod representation,

$$\begin{aligned} &\lim_{\varepsilon\to 0}\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j),\alpha^\varepsilon(t_j))[f(\widehat{\psi}^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - f(\widehat{\psi}^\varepsilon(t),\alpha^\varepsilon(t))] \\ &= \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}(t_j),\alpha(t_j))[f(\widehat{\psi}(t+s),\alpha(t+s)) - f(\widehat{\psi}(t),\alpha(t))]. \end{aligned} \tag{3.50}$$

On the other hand, choose a sequence n_ε such that $n_\varepsilon \to \infty$ as $\varepsilon \to 0$, but $\varepsilon\, n_\varepsilon \to\ 0$. Divide $[t, t+s]$ into intervals of width $\delta_\varepsilon = \varepsilon\, n_\varepsilon$. We have

$$\begin{aligned}
&\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j))[f(\widehat{\psi^\varepsilon}(t+s), \alpha^\varepsilon(t+s)) - f(\widehat{\psi^\varepsilon}(t), \alpha^\varepsilon(t))] \\
&= \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon + n\varepsilon}, \alpha_{ln_\varepsilon + n_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon + n_\varepsilon}, \alpha_{ln_\varepsilon})]\right. \\
&\qquad\qquad \left. + \sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon + n_\varepsilon}, \alpha_{ln_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon}, \alpha_{ln_\varepsilon})]\right].
\end{aligned} \tag{3.51}$$

Since $f(\cdot, \alpha)$ is smooth and bounded, we obtain that

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon + n_\varepsilon}, \alpha_{ln_\varepsilon + n_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon + n_\varepsilon}, \alpha_{ln_\varepsilon})]\right] \\
&= \lim_{\varepsilon\to 0} \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon}, \alpha_{ln_\varepsilon + n_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon}, \alpha_{ln_\varepsilon})]\right].
\end{aligned} \tag{3.52}$$

Thus we need only work with the latter term. Moreover, letting $\varepsilon \to 0$ and $l\delta_\varepsilon = \varepsilon\, ln_\varepsilon \to \varepsilon$ and using nested expectation, we can insert $\boldsymbol{E}_k$ and obtain

$$\begin{aligned}
&\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon}, \alpha_{ln_\varepsilon + n_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon}, \alpha_{ln_\varepsilon})]\right] \\
&= \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \sum_{k=ln_\varepsilon}^{ln_\varepsilon + n_\varepsilon - 1} [f(\widehat{\psi}_{ln_\varepsilon}, \overline{\alpha}_i)\right. \\
&\qquad\qquad \left. \times P(\alpha_{k+1} = \overline{\alpha}_i | \alpha_k = \overline{\alpha}_j) - f(\widehat{\psi}_{ln_\varepsilon}, \overline{\alpha}_j)] I_{\{\alpha_k = \overline{\alpha}_j\}}\right] \\
&= \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon - 1} \left[\frac{\delta_\varepsilon}{n_\varepsilon} \sum_{j=1}^{\infty} \sum_{k=ln_\varepsilon}^{ln_\varepsilon + n_\varepsilon - 1} Qf(\widehat{\psi}_{ln_\varepsilon, \cdot})(\alpha_k) I_{\{\alpha_k = \overline{\alpha}_j\}}\right]\right] \\
&\to \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi^\varepsilon}(t_j), \alpha^\varepsilon(t_j)) \left[\int_t^{t+s} Qf(\widehat{\psi}(u), \alpha(u)) du\right] \quad \text{as} \quad \varepsilon \to 0.
\end{aligned} \tag{3.53}$$

Since $\widehat{\psi}^{\varepsilon}_{ln_\varepsilon}$ and α_{ln_ε} are $\mathcal{F}_{ln_\varepsilon}$-measurable, by virtue of the continuity and boundedness of $\nabla f(\cdot,\alpha)$,

$$\begin{aligned}
&\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^{\varepsilon}(t_j),\alpha^{\varepsilon}(t_j)) \sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1} [f(\widehat{\psi}_{ln_\varepsilon+n_\varepsilon},\alpha_{ln_\varepsilon}) - f(\widehat{\psi}_{ln_\varepsilon},\alpha_{ln_\varepsilon})] \\
&= \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^{\varepsilon}(t_j),\alpha^{\varepsilon}(t_j)) \sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1} \left[\varepsilon\nabla f'(\widehat{\psi}_{ln_\varepsilon},\alpha_{ln_\varepsilon}) \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon}(Y_{k+1}-\widehat{\psi}_k)\right] + o(1).
\end{aligned}$$

where $o(1)\to 0$ as $\varepsilon\to 0$. Next, consider the term

$$\lim_{\varepsilon\to 0}\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^{\varepsilon}(t_j),\alpha^{\varepsilon}(t_j)) \left[\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1} \delta_\varepsilon \left[\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1}\right]\right]. \tag{3.54}$$

Consider a fixed-α process $Y_k(\alpha)$, which is a process with α_k fixed at $\alpha_k=\alpha$ for $ln_\varepsilon \le k \le O(1/\varepsilon)$. Close scrutiny of the inner summation shows that

$$\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1} \quad \text{can be approximated by} \quad \frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1}(\alpha) \tag{3.55}$$

with an approximation error going to 0, since, $\boldsymbol{E}_{ln_\varepsilon}[Y_{k+1}-Y_{k+1}(\alpha)] = O(\eta) = O(\varepsilon)$ by use of the transition matrix (3.2). Thus we have

$$\begin{aligned}
&\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1} \\
&= \sum_{j=1}^{\infty}\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}\left(Y_{k+1}(\overline{\alpha}_j) I_{\{\alpha_{ln_\varepsilon}=\overline{\alpha}_j\}} | \alpha_{ln_\varepsilon}=\overline{\alpha}_j\right) + o(1) \\
&= \sum_{j=1}^{\infty}\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sum_{j_1=1}^{\infty} e_{j_1}[A(\overline{\alpha}_j)]^{k+1-ln_\varepsilon} I_{\{\alpha_{ln_\varepsilon}=\overline{\alpha}_j\}} + o(1),
\end{aligned}$$

where $o(1)\to 0$ in probability as $\varepsilon\to 0$. Henceforth, we write $\mathbb{1}$ in lieu of $\mathbb{1}_\infty$. Note that for

each $j = 1, 2, \ldots$, as $n_\varepsilon \to \infty$ (recall that $\delta_\varepsilon = \varepsilon\, n_\varepsilon$),

$$\frac{1}{n_\varepsilon} \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} [A(\overline{\alpha}_j)]^{k+1-ln_\varepsilon} \to \mathbb{1}\psi'(\overline{\alpha}_j).$$

Note that $I_{\{\alpha_{ln_\varepsilon}=\overline{\alpha}_j\}}$ can be written as $I_{\{\alpha^\varepsilon(l\delta_\varepsilon)=\overline{\alpha}_j\}}$. As $\varepsilon \to 0$ and $l\delta_\varepsilon \to u$, by the weak convergence of $\alpha^\varepsilon(\cdot)$ to $\alpha(\cdot)$ and the Skorohod representation, $I_{\{\alpha^\varepsilon(\varepsilon\, ln_\varepsilon)=\overline{\alpha}_j\}} \to I_{\{\alpha(u)=\overline{\alpha}_j\}}$ w.p.1. Consequently, since $\mathbb{1}\psi'(\overline{\alpha}_j)$ has identical rows,

$$\begin{aligned} \frac{1}{n_\varepsilon} \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1} &\to \sum_{j=1}^{\infty} \psi(\overline{\alpha}_j) I_{\{\alpha(u)=\overline{\alpha}_j\}} \\ &= \psi(\alpha(u)). \end{aligned} \tag{3.56}$$

That is, the limit does not depend on the value of initial state, a salient feature of Markov chains. As a result,

$$\begin{aligned} &\lim_{\varepsilon\to 0} \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1} \frac{1}{n_\varepsilon} \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \boldsymbol{E}_{ln_\varepsilon} Y_{k+1} \right] \\ &= \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{j=1}^{\infty} \int_t^{t+s} \psi(\overline{\alpha}_j) I_{\{\alpha(u)=\overline{\alpha}_j\}} du \right] \\ &= \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j), \alpha^\varepsilon(t_j)) \left[\int_t^{t+s} \psi(\alpha(u)) du \right]. \end{aligned} \tag{3.57}$$

Likewise, it can be shown that, as $\varepsilon \to 0$,

$$\begin{aligned} &\lim_{\varepsilon\to 0} \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1} \delta_\varepsilon \frac{1}{n_\varepsilon} \sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1} \widehat{\psi}_k \right] \\ &= \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(\widehat{\psi}^\varepsilon(t_j), \alpha^\varepsilon(t_j)) \left[\int_t^{t+s} \widehat{\psi}(u) du \right]. \end{aligned} \tag{3.58}$$

Combining (3.50), (3.53), (3.57), and (3.58), the desired result follows. □

3.5 Switching Diffusion Limit

By Theorem 3.1, $\left\{\frac{\widehat{\psi}_n - \boldsymbol{E}\psi(\alpha_n)}{\sqrt{\varepsilon}}\right\}$ is tight for $n \geq\ n_0$, for some positive integer n_0. We define a scaled sequence of the tracking errors $\{v_n\}$ and its continuous-time interpolation $v^\varepsilon(\cdot)$ by

$$v_n = \frac{\widehat{\psi}_n - \boldsymbol{E}\psi(\alpha_n)}{\sqrt{\varepsilon}}, \quad n \geq\ n_0, \quad v^\varepsilon(t) = v_n \quad \text{for} \quad t \in [n\varepsilon, n\varepsilon + \varepsilon) \tag{3.59}$$

to evaluate the rate of variation of the tracking error sequence. Note that from Proposition 3.6,

$$\boldsymbol{E}\psi(\alpha_n) = \overline{\psi}(\varepsilon\ n) + O(\varepsilon), \quad \text{where } \overline{\psi}_n \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} z^i(\varepsilon\ n)\psi(\overline{\alpha}_i), \tag{3.60}$$

where $z^i(t)$ is the ith component of $z(t)$ given in Proposition 3.6. By (M), $\{\alpha_n\}$ is a Markov chain with stationary (time-invariant) transition probabilities, so in view of (3.3),

$$v_{n+1} = v_n - \varepsilon v_n + \sqrt{\varepsilon}(Y_{n+1} - \boldsymbol{E}\psi(\alpha_n)) + \frac{\boldsymbol{E}[\psi(\alpha_n) - \psi(\alpha_{n+1})]}{\sqrt{\varepsilon}}. \tag{3.61}$$

Similarly to the rate of convergence study when α is a fixed parameter (see [90, Chapter 10]), the scaling factor $\sqrt{\varepsilon}$, together with the asymptotic covariance of the limit process, gives us a "rate of convergence" result. However, since v_n is an infinite-dimensional vector, it is not convenient to analysis its limit process. Suppose that $f(\cdot)$ is an arbitrary bounded real-valued function defined on $\mathcal{M}$ such that $\{f(s_j) : 1 \leq\ j < \infty\} \in \ell_1$ $(l = 1)$. Let $y_n = \sum_{j=1}^{\infty} f(s_j)v_n^j$, consider

$$y_{n+1} = y_n - \varepsilon y_n + \sqrt{\varepsilon}\sum_{j=1}^{\infty} f(s_j)(Y_{n+1}^j - \boldsymbol{E}\psi^j(\alpha_n)) + \sum_{j=1}^{\infty} f(s_j)\frac{\boldsymbol{E}[\psi^j(\alpha_n) - \psi^j(\alpha_{n+1})]}{\sqrt{\varepsilon}}. \tag{3.62}$$

We will derive a limit process for $y^\varepsilon(\cdot)$ as $\varepsilon \to 0$. Our task in what follows is to figure out the asymptotic properties of $y^\varepsilon(\cdot)$. Here $y^\varepsilon(t) = y_n$ for $t \in [n\varepsilon, n\varepsilon + \varepsilon)$. We aim to show that the limit is a switching diffusion using a martingale problem formulation.

3.6 Truncation and Tightness

Owing to the definition (3.59), $\{y_n\}$ is not a priori bounded. A convenient way to circumvent this difficulty is to use a truncation device [90]. Let $N > 0$ be a fixed but otherwise arbitrary real number, $S_N(y) = \{y \in \mathbb{R} : |y| \leq N\}$ be the interval with length $2N$, and $\tau^N(y)$ be a smooth function satisfying

$$\tau^N(y) = \begin{cases} 1 & \text{if } |y| \leq N, \\ 0 & \text{if } |y| \geq N+1. \end{cases}$$

Note that $\tau^N(y)$ is "smoothly" connected between the intervals S_N and S_{N+1}. Now define

$$y_{n+1}^N = y_n^N - \varepsilon y_n^N \tau^N(y_n^N) + \sqrt{\varepsilon} \sum_{j=1}^{\infty} f(s_j)(Y_{n+1}^j - \boldsymbol{E}\psi^j(\alpha_n)) + \sum_{j=1}^{\infty} f(s_j) \frac{\boldsymbol{E}[\psi^j(\alpha_n) - \psi^j(\alpha_{n+1})]}{\sqrt{\varepsilon}}. \tag{3.63}$$

and define $y^{\varepsilon,N}(\cdot)$ to be the continuous-time interpolation of y_n^N. It then follows that

$$\lim_{k_0 \to \infty} \limsup_{\varepsilon \to 0} P\left(\sup_{0 \leq t \leq T} |y^{\varepsilon,N}(t)| \geq k_0 \right) = 0 \quad \text{for each} \quad T < \infty$$

and that $y^{\varepsilon,N}(\cdot)$ is a process that is equal to $y^\varepsilon(\cdot)$ up until the first exit from S_N, and hence an N-truncation process of $y^\varepsilon(\cdot)$ [90, p. 284]. To proceed, we work with $\{y^{\varepsilon,N}(\cdot)\}$ and derive its tightness and weak convergence first. Finally, we let $N \to \infty$ to conclude the proof.

Lemma 3.8. *Under conditions* (M) *and* (S), $\{y^{\varepsilon,N}(\cdot)\}$ *is tight in* $D([0, T]; \mathbb{R})$, *and the*

process $\{y^{\varepsilon,N}(\cdot), \alpha^{\varepsilon}(\cdot)\}$ *is tight in* $D([0,\ T];\mathbb{R}\ \times \mathcal{M})$.

Proof. In fact, only the first assertion needs to be verified. In view of (3.63), for any $\delta > 0$ and $t,\ s \leq 0$ with $s \leq \delta$,

$$y^{\varepsilon,N}(t+s) - y^{\varepsilon,N}(t) = -\varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} y_k^N \tau^N(y_k^N) + \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)) + \frac{1}{\sqrt{\varepsilon}} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)\boldsymbol{E}[\psi^j(\alpha_n) - \psi^j(\alpha_{n+1})]. \tag{3.64}$$

Owing to the N-truncation used,

$$\boldsymbol{E}_t^{\varepsilon} \left| \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} y_k^N \tau^N(y_k^N) \right|^2 \leq Ks,$$

and as a result,

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} \boldsymbol{E} \sup_{0\leq s\leq \delta} \boldsymbol{E}_t^{\varepsilon} \left| \varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} y_k^N \tau^N(y_k^N) \right|^2 = 0 \tag{3.65}$$

Next, by virtue of (M), the irreducibility of the conditional Markov chain $\{Y_n\}$ implies that it is ϕ-mixing with exponential mixing rate [76, p. 167], $\boldsymbol{E}\psi(\alpha_k) - \boldsymbol{E}Y_{k+1} \to 0$ exponentially

fast, and consequently

$$\begin{aligned}
&\boldsymbol{E}_t^\varepsilon \left| \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)) \right|^2 \\
&= \boldsymbol{E}_t^\varepsilon \left| \sqrt{\varepsilon} \sum_{j=1}^{\infty} \left(f(s_j) \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} (Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)) \right) \right|^2 \\
&= \boldsymbol{E}_t^\varepsilon \left| \sqrt{\varepsilon} \sum_{j=1}^{\infty} \left(f(s_j) \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} [(Y_{k+1}^j - \boldsymbol{E}Y_{k+1}^j) - (\boldsymbol{E}\psi^j(\alpha_k) - \boldsymbol{E}Y_{k+1}^j)] \right) \right|^2 = O(s).
\end{aligned}$$

This yields that

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} \boldsymbol{E} \sup_{0\le s\le \delta} \boldsymbol{E}_t^\varepsilon \left| \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)) \right|^2 = 0. \tag{3.66}$$

In addition,

$$\begin{aligned}
\frac{1}{\sqrt{\varepsilon}} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)\boldsymbol{E}[\psi^j(\alpha_n) - \psi^j(\alpha_{n+1})] &= \sum_{j=1}^{\infty} \left(f(s_j) \frac{1}{\sqrt{\varepsilon}} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \boldsymbol{E}[\psi^j(\alpha_n) - \psi^j(\alpha_{n+1})] \right) \\
&= \sum_{j=1}^{\infty} \left(f(s_j) \frac{1}{\sqrt{\varepsilon}} [\boldsymbol{E}\psi^j(\alpha_{t/\varepsilon}) - \boldsymbol{E}\psi^j(\alpha_{(t+s)/\varepsilon})] \right) \\
&= O(\sqrt{\varepsilon})
\end{aligned} \tag{3.67}$$

Combining (3.64)-(3.67), we have

$$\lim_{\delta\to 0} \limsup_{\varepsilon\to 0} \boldsymbol{E} \left\{ \sup_{0\le s\le \delta} \boldsymbol{E}_t^\varepsilon |y^{\varepsilon,N}(t\ +s) - y^{\varepsilon,N}(t)|^2 \right\} = 0,$$

and hence the criterion [77, p. 47] implies that $\{y^{\varepsilon,N}(\cdot)\}$ is tight. □

3.7 Representation of Covariance

The main results to follow, Lemma 3.9 and Corollary 3.10 for the diffusion limit in Section 3.7.1, require representation of the covariance of the conditional Markov chain $\{Y_k\}$. This is again worked out via the use of fixed-α process $Y_k(\alpha)$ similar in spirit to (3.55). For any integer $m \geq 0$, for $m \leq k \leq O(1/\varepsilon)$, with α_k fixed at α, $Y_{k+1}(\alpha)$ is a Markov chain with 1-step irreducible transition matrix $A(\alpha)$ and stationary distribution $\psi(\alpha)$. Thus [76, p. 167] implies that $\{Y_{k+1}(\alpha) - \boldsymbol{E}Y_{k+1}(\alpha)\}$ is a ϕ-mixing sequence with zero mean and exponential mixing rate, and hence it is strongly ergodic. Similarly to (3.55), $Y_{k+1} - \boldsymbol{E}Y_{k+1}$ can be approximated by a fixed α process $Y_{k+1}(\alpha) - \boldsymbol{E}Y_{k+1}(\alpha)$. Taking $n = n_\varepsilon \leq O(1/\varepsilon)$ as $\varepsilon \to 0$, $n \to \infty$, and

$$\begin{aligned}\lim_{\varepsilon\to 0}\frac{1}{n}\sum_{k_1=m}^{n+m-1}\sum_{k=m}^{n+m-1}\sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^{j}(\alpha)-\boldsymbol{E}Y_{k+1}^{j}(\alpha))\sum_{j_1=1}^{\infty} f(s_{j_1})(Y_{k_1+1}^{j_1}(\alpha)-\boldsymbol{E}Y_{k_1+1}^{j_1}(\alpha))\\ = \sigma(\alpha) \quad \text{w.p.1},\end{aligned} \tag{3.68}$$

and

$$\lim_{\varepsilon\to 0}\frac{1}{n}\sum_{k_1=m}^{n+m-1}\sum_{k=m}^{n+m-1}\boldsymbol{E}\sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^{j}(\alpha)-\boldsymbol{E}Y_{k+1}^{j}(\alpha))\sum_{j_1=1}^{\infty} f(s_{j_1})(Y_{k_1+1}^{j_1}(\alpha)-\boldsymbol{E}Y_{k_1+1}^{j_1}(\alpha)) = \sigma(\alpha) \tag{3.69}$$

Note that (3.68) is a consequence of ϕ-mixing and strong ergodicity, and (3.69) follows from (3.68) by means of the dominated convergence theorem. By [93, Theorem 2.13], moreover detailed computation yields an explicit formula for the limit variance

$$\sigma^2(i) = \sum_{\imath=1}^{\infty}\sum_{\jmath=1}^{\infty} f(s_\imath)f(s_\jmath)\left[\boldsymbol{E}\psi^\imath\sum_{k=0}^{\infty} v_0^{\imath\jmath}(k) + \boldsymbol{E}\psi^\jmath\sum_{k=0}^{\infty} v_0^{\jmath}(k)\right], \tag{3.70}$$

and $v_0^{\imath\jmath}(k)$ is the $\imath\jmath$th entry of $V_0(k)$ given by

$$V_0(k) = \left(\begin{array}{c} \\ I \ - \\ \\ \end{array} \left(\begin{array}{c} \boldsymbol{E}\psi \\ \boldsymbol{E}\psi \\ \vdots \\ \end{array} \right) \right)$$

3.7.1 Weak Limit as Solution of a Martingale Problem

To obtain the desired weak convergence result, we work with the pair $(y^{\varepsilon,N}(\cdot), \alpha^{\varepsilon}(\cdot))$. By virtue of the tightness and Prohorov's theorem, we can extract a weakly convergent subsequence (still denoted by $(y^{\varepsilon,N}(\cdot), \alpha^{\varepsilon}(\cdot))$ for simplicity) with limit $(y^N(\cdot),\ \alpha(\cdot))$. We will show that the limit is a switching diffusion.

To proceed with the diffusion approximation, similarly as in the proof of Theorem 3.7, we will use the martingale problem formulation to derive the desired result. For $y \in \mathbb{R}$, $\alpha \in \mathcal{M}$, and any twice continuously differentiable function $g(\cdot, \alpha)$ with compact support, consider the operator $\mathcal{L}$ defined by

$$\mathcal{L}g(y,i) = -\frac{\partial}{\partial y} g(y,i)y + \frac{1}{2}\sigma^2(i)\frac{\partial^2}{\partial y^2} g(y,i) + Qg(y,\cdot)(i), \quad i \in \mathcal{M} \tag{3.71}$$

where $\sigma^2(i) > 0$ will be specified later. We will show that the limit process is a solution of a martingale problem with operator $\mathcal{L}$, which has a unique solution. As a result, the limit is a switching diffusion process. For any positive integer ℓ_0, any $t > 0$, $s > 0$, any $0 < t_j \leq t$ with $j \leq \ell_0$, and any bounded and continuous function $h_j((\cdot),\ \alpha)$ for each $\alpha \in \mathcal{M}$, we aim to derive an equation similar to (3.49) with the operator L_1 replaced by $\mathcal{L}$. As in the proof of Theorem 3.7, we work with the sequence indexed by ε. Choose $n\varepsilon$ such that $n\varepsilon \to \infty$ but

$\delta_\varepsilon = \varepsilon\, n_\varepsilon \to 0$. The tightness of $\{y^{\varepsilon,N}(\cdot), \alpha^\varepsilon(\cdot)\}$ and the Skorohod representation yield that (3.50)-(3.52) hold with $\widehat{\psi}^\varepsilon(\cdot)$ and $\widehat{\psi}(\cdot)$ replaced by $y^{\varepsilon,N}(\cdot)$ and $y^N(\cdot)$, respectively.

Lemma 3.9. *Assume the conditions of Lemma* 3.8 *and that* $(y^{\varepsilon,N}(0),\ \alpha^\varepsilon(0))$ *converges weakly to* $(y^N(0),\ \alpha(0))$. *Then* $(y^{\varepsilon,N}(\cdot),\ \alpha^\varepsilon(\cdot))$ *converges weakly to* $(y^N(\cdot),\ \alpha(\cdot))$, *which is a solution of the martingale problem with operator* $\mathcal{L}^N$ *given by*

$$\mathcal{L}^{\mathcal{N}} g(y,\alpha) = -\frac{\partial}{\partial y} g(y^N,\alpha) y^N \tau^N(y^N) + \frac{1}{2}\sigma^2(\alpha)\frac{\partial^2}{\partial y^2} g(y^N,\alpha) + Qg(y^N,\cdot)(\alpha), \qquad \alpha \in \mathcal{M} \tag{3.72}$$

Proof. In view of (3.67), the term $\sum_{k=t/\varepsilon}^{(t+s)/\varepsilon} \sum_{j=1}^{\infty} f(s_j)[\boldsymbol{E}\psi^j(\alpha_k) - \boldsymbol{E}\psi^j(\alpha_{k+1})]/\sqrt{\varepsilon} = O(\sqrt{\varepsilon})$ can be ignored in the characterization of the limit process. Moreover,

$$\begin{aligned}
&\sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)[Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)] \\
&= \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^j - \boldsymbol{E}Y_{k+1}^j) + \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(\boldsymbol{E}Y_{k+1}^j - \boldsymbol{E}\psi^j(\alpha_k)).
\end{aligned}$$

Since $\boldsymbol{E}Y_{k+1} - \boldsymbol{E}\psi(\alpha_k) \to 0$ exponentially fast owing to the elementary properties of a Markov chain, the last term above is $o(1)$ that goes to 0 as $\varepsilon \to 0$. Thus,

$$y^{\varepsilon,N}(t+s) - y^{\varepsilon,N}(t) = -\varepsilon \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} y_k^N \tau^N(y_k^N) + \sqrt{\varepsilon} \sum_{k=t/\varepsilon}^{(t+s)/\varepsilon-1} \sum_{j=1}^{\infty} f(s_j)(Y_{k+1}^j - \boldsymbol{E}Y_{k+1}^j) + o(1). \tag{3.73}$$

Similarly to the argument in the proof of Theorem 3.7,

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j), \alpha^\varepsilon(t_j)) \left[\sum_{ln_\varepsilon = t/\varepsilon}^{(t+s)/\varepsilon-1} [g(y^N_{ln_\varepsilon, \alpha_{ln_\varepsilon+n_\varepsilon}}) - g(y^N_{ln_\varepsilon}, \alpha_{ln_\varepsilon})] \right] \\
&\quad = \boldsymbol{E} \prod_{j=1}^{\ell_0} h_j(y^N(t_j), \alpha(t_j)) \left[\int_t^{t+s} Qg(y^N(u), \alpha(u)) du \right].
\end{aligned} \tag{3.74}$$

In addition,

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^{\varepsilon}(t_j))\left[-\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\delta_\varepsilon}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\frac{\partial}{\partial\, y}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})y^N_k\tau^N(y^N_k)\right]\\
&\quad=\lim_{\varepsilon\to 0} \boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^{\varepsilon}(t_j))\left[-\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\delta_\varepsilon\frac{\partial}{\partial\, y}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})y^N_{ln_\varepsilon}\tau^N(y^N_{ln_\varepsilon})\right] \qquad (3.75)\\
&\quad=\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{N}(t_j),\alpha(t_j))\left[-\int_t^{t+s}\frac{\partial}{\partial\, y}g(y^N(u),\alpha(u))y^N(u)\tau^N(y^N(u))du\right].
\end{aligned}$$

Next we note that

$$\begin{aligned}
&\left|\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^{\varepsilon}(t_j))\left[\sqrt{\varepsilon}\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\partial}{\partial\, y}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sum_{\jmath=1}^{\infty}f(s_\jmath)[Y^\jmath_{k+1}-\boldsymbol{E}\ Y^\jmath_{k+1}]\right]\right|\\
&\quad\le\left|\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^{\varepsilon}(t_j))\left[\sqrt{\varepsilon}\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\left|\frac{\partial}{\partial\, y}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})\right|\right.\right.\\
&\qquad\qquad\left.\left.\times\sum_{\jmath=1}^{\infty}f(s_\jmath)\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}|\boldsymbol{E}_{ln_\varepsilon}[Y^\jmath_{k+1}-\boldsymbol{E}\ Y^\jmath_{k+1}]|\right]\right|\\
&\quad\to 0\quad\text{as}\quad\varepsilon\to\ 0
\end{aligned}\qquad (3.76)$$

owing to the mixing property.

Finally, define

$$g^2_{ln_\varepsilon}=\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\boldsymbol{E}_{ln_\varepsilon}\left(\sum_{j=1}^{\infty}f(s_j)[Y^j_{k+1}-\boldsymbol{E}\ Y^j_{k+1}]\sum_{j_1=1}^{\infty}f(s_{j_1})[Y^{j_1}_{k+1}-\boldsymbol{E}\ Y^{j_1}_{k+1}]\right)$$

It follows that

$$\begin{aligned}
&\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^\varepsilon(t_j))\left[\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\partial^2}{\partial y^2}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})(y^N_{ln_\varepsilon+n_\varepsilon}-y^N_{ln_\varepsilon})^2\right]\\
&=\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^\varepsilon(t_j))\left[\sum_{j=1}^{\infty}\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\partial^2}{\partial y^2}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})(y^N_{ln_\varepsilon+n_\varepsilon}-y^N_{ln_\varepsilon})^2 I_{\{\alpha_{ln_\varepsilon}=\overline{\alpha}_j\}}\right]\\
&=\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^\varepsilon(t_j))\left[\sum_{j=1}^{\infty}\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\delta_\varepsilon\frac{\partial^2}{\partial y^2}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})\boldsymbol{E}_{ln_\varepsilon}g^2_{ln_\varepsilon}I_{\{\alpha_{ln_\varepsilon}=\overline{\alpha}_j\}}\right]+\rho_\varepsilon,
\end{aligned}$$

where $\rho_\varepsilon \to 0$ as $\varepsilon \to 0$. Since it is conditioned on $\alpha_{ln_\varepsilon} = \alpha_j$, $Y_{k+1} - \boldsymbol{E}\, Y_{k+1}$ can be approximated by a fixed-$\overline{\alpha}_j$ process $Y_{k+1}(\alpha_j) - \boldsymbol{E}Y_{k+1}(\alpha_j)$, and since $Y_{k+1}(\alpha_j) - \boldsymbol{E}Y_{k+1}(\alpha_j)$ is a Markov chain with irreducible transition matrix $A(\alpha_j)$, it is ϕ-mixing, and the argument in (3.69) implies that for each $\alpha_j \in \mathcal{M}$ with $j = 1, 2, \ldots,$

$$\begin{aligned}
&\frac{1}{n_\varepsilon}\sum_{k=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\sum_{k_1=ln_\varepsilon}^{ln_\varepsilon+n_\varepsilon-1}\boldsymbol{E}_{ln_\varepsilon}\left(\sum_{j=1}^{\infty}f(s_j)[Y^j_{k+1}-\boldsymbol{E}Y^j_{k+1}]\sum_{j_1=1}^{\infty}f(s_{j_1})[Y^{j_1}_{k+1}-\boldsymbol{E}Y^{j_1}_{k+1}]\right)\\
&\quad\to\sigma(\overline{\alpha}_j)\quad\text{w.p.1 as}\quad\varepsilon\to 0,
\end{aligned}\tag{3.77}$$

where $\sigma(\alpha)$ is defined in (3.69). By virtue of Lemma 3.5, $\alpha^\varepsilon(\cdot)$ converges weakly to $\alpha(\cdot)$. As a result, by Skorohod representation, sending $\varepsilon \to 0$ and $l\delta_\varepsilon \to u$ leads to $\alpha^\varepsilon(\varepsilon\, ln_\varepsilon)$ converging to $\alpha(u)$ w.p.1. In addition, $I_{\{\alpha^\varepsilon(l\delta_\varepsilon)=\overline{\alpha}_j\}} \to I_{\{\alpha(u)=\overline{\alpha}_j\}}$ w.p.1. It follows that

$$\begin{aligned}
&\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{\varepsilon,N}(t_j),\alpha^\varepsilon(t_j))\left[\sum_{ln_\varepsilon=t/\varepsilon}^{(t+s)/\varepsilon-1}\frac{\partial^2}{\partial y^2}g(y^N_{ln_\varepsilon},\alpha_{ln_\varepsilon})(y^N_{ln_\varepsilon+n_\varepsilon}-y^N_{ln_\varepsilon})^2\right]\\
&\to\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{N}(t_j),\alpha(t_j))\left[\int_t^{t+s}\sum_{j=1}^{\infty}[\frac{\partial^2}{\partial y^2}g(y^N(u),\overline{\alpha}_j)\sigma(\overline{\alpha}_j)]I_{\{\alpha(u)=\overline{\alpha}_j\}}du\right]\\
&=\boldsymbol{E}\prod_{j=1}^{\ell_0} h_j(y^{N}(t_j),\alpha(t_j))\left[\int_t^{t+s}[\frac{\partial^2}{\partial y^2}g(y^N(u),\alpha(u))\sigma(\alpha(u))]du\right]
\end{aligned}\tag{3.78}$$

In view of (3.74)-(3.78), the desired result follows. □

Corollary 3.10. *Under the conditions of Lemma* 3.9*, the untruncated process* $(y^\varepsilon(\cdot),\ \alpha^\varepsilon(\cdot))$ *converges weakly to* $(y(\cdot), \alpha(\cdot))$ *satisfying the switching diffusion equation*

$$dy(t) \ = -y(t)dt \ + \ \sigma(\alpha(t))dB, \tag{3.79}$$

where $B(\cdot)$ *is a 1-dimensional standard Brownian motion and* $\sigma(\alpha)$ *is given by* (3.69)*.*

Proof. The uniqueness of the associated martingale problem can be proved similarly to that of [80, Lemma 7.18]. The rest of the proof follows from a similar argument as in [90, Step 4, p. 285]. □

3.8 An application on adaptive discrete stochastic optimization

Consider the following discrete stochastic optimization problem:

$$\min_{\overline{\alpha}\in\mathcal{M}} \boldsymbol{E}\{c_n(\overline{\alpha})\}, \tag{3.80}$$

where for each fixed $\overline{\alpha} \in \mathcal{M}$, $\{c_n(\alpha)\}$ is a sequence of i.i.d. random variables with finite variance. we assume that the $\mathcal{M}$ in (3.1) is $\mathcal{M} \ = \ \mathcal{S} \ = \ \{e_1,\ e_2,\ \ldots,\ \}$, where e_i denotes the standard unit vector with infinite dimensions. In what follows, $\mathcal{M}$ denotes the set of candidate values from which the time-varying global minimizer is chosen at each time instant (according to a slow Markov chain). $\mathcal{S}$ is the set of candidate solutions for the discrete optimization. Because we assume $\mathcal{M} = \ \mathcal{S}$, we do not use the notation $\mathcal{S}$ in this section.

Remark 3.11. Let $\mathcal{K} \subset \mathcal{M}$ denote the set of global minimizers for (3.80). If the global minima set $\mathcal{K}$ does not evolve with time, we say the problem is static. The situation has been discussed in [94] and [95]. In [92], the authors discussed the case that $\mathcal{M} = \mathcal{S} = \{e_1, \ldots, e_S\}$

have finite states and the global minima set $\mathcal{K}$ of (3.80) is time varying. Here we will consider the situation that $\mathcal{M} = \mathcal{S}$ have countable states. So this is an extended version of a discrete stochastic optimization algorithm proposed by Andradóttir [94].

The following stochastic ordering assumptions are used.

(O) For each e_i, $e_j \in \mathcal{M}$, there exists some random variable Z^{e_i,e_j} such that for all $e_i \in \mathcal{K}$, $e_j \in \mathcal{K}$, and $e_l \in \mathcal{M}$, $l \neq i,\ j$,

$$\begin{aligned} &P(Z^{e_j,e_i} > 0) \geq P(Z^{e_i,e_j} > 0), \quad P(Z^{e_l,e_i} > 0) \geq P(Z^{e_l,e_j} > 0), \\ &P(Z^{e^i,e^l} \leq 0) \geq P(Z^{e_j,e_l} \leq 0). \end{aligned} \tag{3.81}$$

Denote this time-varying optimal solution as α_n. We subsequently refer to α_n as the *true parameter* or *hypermodel*. Tracking such time-varying parameters is at the very heart of applications of adaptive SA algorithms. The adaptive algorithm are proposed as follows.

Algorithm 1. (adaptive discrete stochastic optimization algorithm)

`Step 0`: (Initialization) At time $n = 0$, select starting point $Y_0 \in \mathcal{M}$. Set $\widehat{\psi}_0 = Y_0$, and select $\widehat{\alpha}_0^* = Y_0$.

`Step 1`: (Random search) At time n, sample $\widetilde{Y}_n$ with uniform distribution from $\mathcal{M} - \{Y_n\}$.

`Step 2`: (Evaluation and acceptance) Generate observation $Z^{Y_n,\widetilde{Y}_n}$. If $Z^{Y_n,\widetilde{Y}_n} > 0$, set $Y_{n+1} = \widetilde{Y}_n$; else, set $Y_{n+1} = Y_n$.

`Step 3`: (LMS algorithm for updating occupation probabilities of Y_n) Construct $\widehat{\psi}_{n+1}$ as

$$\widehat{\psi}_{n+1} = \widehat{\psi}_n + \varepsilon(Y_{n+1} - \widehat{\psi}_n). \tag{3.82}$$

`Step 4`: (Compute estimate of the solution) $\widehat{\alpha}_n^* = e_{i^*}$, where

$$i^* = \arg \max_{i \in \{1,2,\ldots\}} \widehat{\psi}_{n+1}^*;$$

set $n \to n+1$ and go to Step 1 ($\widehat{\psi}^i_{n+1}$ denotes the ith component of the vector $\widehat{\psi}_{n+1}$).

Note that as long as $0 < \varepsilon < 1$, $\widehat{\psi}_n$ is guaranteed to be a probability vector. Intuitively, the constant step size ε introduces exponential forgetting of the past occupation probabilities and permits tracking of slowly time-varying α_n. Since $\alpha_n \in \mathcal{M}$ and $\mathcal{M}$ is a finite state space, it is reasonable to describe $\{\alpha_n\}$ as a slow Markov chain on $\mathcal{M}$ for the subsequent analysis. Henceforth, we assume that (M) holds for $\{\alpha_n\}$. Note that the hypermodel assumption is used only for the analysis and does not enter the actual algorithm implementation; see Algorithm 1. By [92, Theorem 6.1], we know that for fixed $\alpha_n = \alpha$ the sequence $\{Y_n\}$ generated by Algorithm 1 is a conditional Markov chain (conditioned on α_n); i.e., assumption (S) of Section 2 holds. The update of the occupation probabilities (3.82) is identical to (3.3). Thus the behavior of the sequence $\{\widehat{\psi}_n\}$ generated by Algorithm 1 exactly fits the model of Section 2. In particular, the mean squares analysis and the limit system of switching ODEs of Section 3, and switching diffusion limit of Section 4 hold.

3.9 Further remarks

This chapter has been devoted to a class of stochastic approximation problems with regime switching modulated by discrete-time Markov chain. Under simple conditions, it has been shown that a continuous-time interpolation of the iterates converges weakly to a system of ODEs with regime switching and that a suitably scaled sequence of the tracking errors converges to a system of switching diffusion. For future study, a worthwhile effort is to

examine Markov decision processes having general state spaces with emphasis on switching diffusion type Markov decision processes. Another direction of considerable interest is to pursue the study of semi-Markov processes.

CHAPTER 4 Asynchronous Stochastic Approximation Algorithms for Networked Systems: Regime-Switching Topologies and Multi-scale Structure

4.1 Introduction

This chapter develops consensus algorithms under the asynchronous communication and random computation environment using random switching topologies. Consensus problems are related to many control applications that involve coordination of multiple entities with only limited neighborhood information to reach a global goal for the entire team. Since the mid 1990s, there have been increasing and resurgent efforts devoted to the study of consensus controls of multi-agent systems. The goal is to achieve a common theme such as position, speed, load distribution, etc. for the mobile agents. In [123], a discrete-time model of autonomous agents was proposed, which can be viewed as points or particles moving in the plane with the same speed but with different headings. Each agent updates its heading using a local rule based on the average headings of its own and its neighbors. This is in fact a special version of a model introduced in [119] for simulating animation of flocking and schooling behaviors. Technically, the problems considered are related to the parallel computation model considered in [121], which was substantially generalized in [107]; see also related works in [96, 99, 101, 111–114, 117, 130]. During the past decades, a host of researchers have devoted their efforts to the study of the consensus problems; see [102–105, 110, 115, 116, 118, 120, 124], and many references therein. Many results obtained thus far are for simple dynamic systems with fixed or highly simplified time-varying topologies, whereas [105], [128], and [129] dealt with time-varying topologies under Markovian switching.

In practical implementations of consensus or coordinated control schemes, control actions

are almost always done asynchronously, especially over a large network of subsystems. For instance, subsystems operate independently with different clocks until they communicate with their neighbors; communication channels operate according to priorities and hence transmit data at different pace and at different time; even for data packets transmitted at the same time from a node system, their pathways through different routes and hubs introduce different latencies and hence arrive at different time. This is especially true in mobile agents when obstacles from terrains create interruptions, packet losses, and delays so that consensus must be performed on delayed information which is an asynchronous operation.

In this chapter, our problems are formulated to capture two aspects of the asynchronism: (i) Asynchronous execution of state updates at the subsystems: Each subsystem has a randomized timer, representing the internal processing time. A subsystem can update its state only when the timer ticks. After the state update, the timer is renewed and internal processing resumes until the next ticking time. (ii) Asynchronous neighborhood information exchange: When a subsystem's timer ticks, the subsystem will observe the states of its neighboring subsystems at that time and adjust its own state accordingly. Since the neighboring subsystems update their states independently, the received state information will always be a delayed information, creating another layer of asynchronism. These concepts are illustrated in Figure 9.

This asynchronous framework introduces fundamental challenges to constrained consensus control problems. In the field of consensus control, most works are on unconstrained consensus, namely, as long as the states of the subsystems achieve consensus there are no other constraints to be satisfied. However, practical systems often impose constraints on the states. For example, for power grids, all power produced will have to be equal to the total load at steady state, even though transient power imbalance is allowed due to storage capabilities

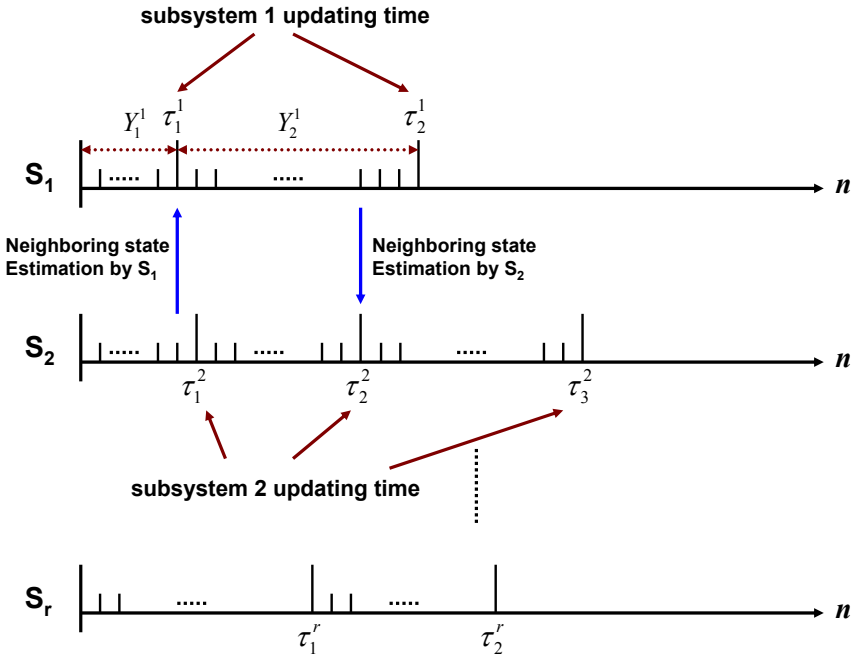


Figure 9: Asynchronous operations and communications of networked subsystems.

on generators and/or capacitance and inductance on transmission lines. In a team formation for area surveillance, a team of mobile sensors need to be confined to the region to be covered. In parallel or cloud computing, steady-state service demands and service capacity must be equal. In our previous work, this constraint is satisfied by employing a "link control" strategy in which a reduction on a state value is always balanced by an equal amount of increase in its neighboring subsystems. The constrained consensus control problems are motivated by load sharing and resource allocation problems. When node i estimates the state of node j and decides to shift a resource of amount u^{ij} to node j, this will be a reduction on node i and an increase of the equal amount to node j. In this sense, both node i and node j are controlled. But the decision resides with node i, and node j receives it passively. In synchronized operation, this will guarantee that the sum of all states is a constant at all time. In asynchronous modes, state updates occur at different times, and as a result, the sum of the

states may not be a constant during the transient period. Asynchronous operation renders such a control scheme impractical. Further complication stems from the time-varying nature of network topologies which changes a subsystem's neighbors randomly. Consequently, the interaction between the stochastic processes of subsystem timers and the governing Markov chain for the network topologies must be carefully studied. In this work, we employ a new control strategy in which the state constraint can be asymptotically satisfied even though the asynchronous operation leaves the constraint unmet during transient.

Dealing with large interconnected systems such as in a communication network with multiple servers, it is natural to consider the distributed, asynchronous stochastic approximation (SA) algorithms. If synchronous SA algorithms are used, a new iteration will not begin until the current iteration is finished in all subsystems. Since the dimension of a networked system can be very large, the waiting time on subsystems will cause serious time delays. In [121], an asynchronous algorithm was proposed where separate processors iterated on the same system vector (with possibly different noise processes and/or different dynamics) and shared information in an asynchronous way. In [107, 108, 135], SA algorithms for parallel and distributed processing were further developed. The main efforts were on the study of convergence and rates of convergence of such algorithms.

In this chapter, we concentrate on consensus-type algorithms. Here, each component in a system (with a large number of mobile agents) can be handled by different agents and the information can be shared by agents. To each single agent, it can start the next iteration using the newest information of iteration on other components without waiting for other agents to finish. So for each component, the time of each iteration and the number of iterations up to that moment are random. We note that the underlying problems introduce some new challenges, and our solutions carry a number of new features. When representing the

algorithms as discrete-time dynamic systems, the system dynamics switch randomly among a finite number of regimes and at random times. The modulating force of the switching process is modeled as a discrete-time Markov chain with a finite-state space. In our setup, the transition probability matrix of the Markov chain includes a small parameter ε. Henceforth, this parameter will be called the transition frequency parameter since it represents how frequently the state transition will take place. On the other hand, the SA algorithm defines its updating speed by another small parameter μ, which will be called the adaptation step-size. The interplay of the two parameters introduces a multi-scale system dynamics. It turns out that the difference between the parameters ($\varepsilon = O(\mu)$, $\varepsilon \ll \mu$, and $\mu \ll \varepsilon$) gives rise to qualitatively different behaviors with stark contrasts.

To summarize, there are several novel features of the algorithms proposed in this chapter. (1) In contrast to the most existing consensus algorithms, the participating agents compute and communicate in an asynchronous fashion. (2) Based on their local clocks, the agents compute and communicate at random times without using a global clock. (3) The regime-switching process is modeled as a discrete-time Markov chain with a finite state space. (4) The functions involved are allowed to vary with respect to time hence nonstationarity can be handled. (5) Multi-scale formulation enriches the applicability of the algorithms.

The rest of the chapter is arranged as follows. Section 4.2 begins with the basic knowledge of consensus. Section 4.3 introduces the formulation of a typical consensus control problem for networked systems under randomly switching topologies. It serves to demonstrate how this problem naturally leads to asynchronous SA algorithms under switching dynamics. Mathematics formulation of the problem is then presented accordingly. Section 4.4 focuses on the case $\varepsilon = O(\mu)$ to introduce new techniques in establishing asymptotic behavior of the algorithms. Using weak convergence methods, convergence of the algorithm is obtained. The

limit behavior of the scaled estimation errors is also analyzed. Section 4.6 extends the main techniques of Section 4.4 to the cases of $\varepsilon \ll \mu$ and $\mu \ll \varepsilon$. It is shown that depending on relative scales between the transition frequency and adaptation step-size, the asynchronous SA algorithms demonstrate fundamentally different asymptotic behaviors. Section 4.7 illustrates the main findings of this chapter by simulation examples. Section 4.8 provides further remarks and discusses some open issues.

4.2 Consensus Algorithm Basics: Traditional Setting

This section gives a brief account on the setup of consensus under simple conditions. Consider a networked system of r nodes, given by

$$x_{n+1}^i = x_n^i + u_n^i, \quad i = 1, \ldots, r, \tag{4.1}$$

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_n^{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 $\mathcal{G}$.

We consider link controls 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.,

$$\widehat{x}_n^{ij} = x_n^j + d_n^{ij}. \tag{4.2}$$

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

$$d_n^{ij} = g(x_n^j, \xi_n^{ij}) \tag{4.3}$$

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.

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 [101, 113]. 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, (4.2) 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 [98] for basic matrix properties in graphs and to [122] 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 $W = H'G$.

Under the link-based state control u_n^i, the state updating scheme (4.1) becomes

$$x_{n+1} = x_n - \mu H' G \delta_n. \tag{4.4}$$

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 (4.4), which for the analysis is

$$x_{n+1} = x_n + \mu(Mx_n + W\xi_n). \tag{4.5}$$

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

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

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. Assume that the noise is 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. The proof of the following theorem is in [128].

Theorem 4.1.. *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.*

4.3 Formulation

Throughout this chapter, $|\cdot|$ denotes a Euclidean norm. A point x in a Euclidean space is a column vector; the ith component of x is denoted by x^i, $\mathbb{1}$ denotes the column vector with all components being 1. The symbol $'$ denotes transpose. The notation $O(y)$ denotes a function of y satisfying $\sup_y |O(y)|/|y| < \infty$. Likewise, $o(y)$ denotes a function of y satisfying $|o(y)|/|y| \to 0$, as $y \to 0$. In particular, $O(1)$ denotes the boundedness and $o(1)$ indicates convergence to 0. To facilitate the reading, we have placed some basic formulation for consensus control algorithms in Section 4.2. Our formulation in this chapter is much beyond the traditional setup. In lieu of the simple formulation in Section 4.2, we allow certain nonadditive noises be added. More importantly, our main effort is on asynchronous computation and communication schemes. In lieu of the constraint $\mathbb{1}'x_n = \eta\, r$ at each step, we only require such an equality to hold asymptotically. This generalizes the setup in Section 4.2 of this chapter. Suppose that the network topology is represented by a graph $\mathcal{G}$. Different from the standard setting, the graph depends on a discrete-time Markov chain so it is given by $\mathcal{G} = \mathcal{G}(\widetilde{\alpha}_n)$. 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, rerouting of communication channels, etc. Thus $\mathcal{G}(\widetilde{\alpha}_n) = \sum_{\iota=1}^{m_0} \mathcal{G}(\iota) I_{\{\widetilde{\alpha}_n = \iota\}}$. To illustrate, suppose that initially the Markov chain is at $\widetilde{\alpha}_0 = \iota$. Then the graph takes the value $\mathcal{G}(\iota)$. At a random instance ρ_1, the first jump of the Markov chain takes place so that $\widetilde{\alpha}_{\rho_1} = \ell \neq \iota$. Then the graph switches to $\mathcal{G}(\ell)$ and holds that value for a random duration until the next jump of the Markov chain takes place etc.

To carry out the recursive computational task, we consider a class of asynchronous and distributed algorithms in the following setup. Suppose that the state $x \in \mathbb{R}^r$ and there are r

processors participating in the computational task. For notational simplicity, we assume that each processor handles only one component. It is clear that this can be made substantially more general by allowing each processor handles a vector of possibly different dimensions. However, the mathematical framework will be essentially the same albeit the complex notation. Suppose that for each $i \leq r$, $\{Y_n^i\}$ is a sequence of positive integer-valued random variables (assuming the random sequence to be positive integer valued is for notational convenience) that are generally state and data dependent such that the nth iteration of processor i takes Y_{n-1}^i units of time. Define a sequence of "renewal-type" random computation times τ_n^i as

$$\tau_0^i = 0, \qquad \tau_{n+1}^i = \tau_n^i + Y_n^i. \tag{4.6}$$

For each i, the sequence $\{Y_n^i\}$ is an inter-arrival time and $\{\tau_n^i\}$ is the corresponding "renewal" time. It is well known that $\widetilde{\alpha}_n$ is strongly Markov, so $\widetilde{\alpha}_{\tau_n^i}$ is a Markov chain.

Using constant stepsize $\mu > 0$, we consider the following asynchronous algorithm

$$x^i_{\tau^i_{n+1}} = x^i_{\tau^i_n} + \mu[M_{\tau^i_n}(\widetilde{\alpha}_{\tau^i_n})x_{\tau^i_n}]^i + \mu[W_{\tau^i_n}(\widetilde{\alpha}_{\tau^i_n})\widetilde{\xi}^i_n]^i + \mu\widehat{W}^i_{\tau^i_n}(x_{\tau^i_n}, \widetilde{\alpha}_{\tau^i_n}, \widetilde{\zeta}^i_n), \quad i \leq r, \tag{4.7}$$

where $\widetilde{\xi}_n^i \in \mathbb{R}^r$ and $\widetilde{\zeta}_n^i \in \mathbb{R}^r$ are the noise sequences incurred in the $(n+1)$st iteration, Note that the functions involved are time dependent. We use the same idea as in the setup of a fixed configuration as in Section 4.2, but allow more general structure. Note also that for each n and $\alpha \in \mathcal{M}$, $M_n(\alpha)$ is not a generator of a Markov chain as the fixed M discussed in Section 4.2. We allow the non-additive noise be used. When $M_n(\iota) = M$ and $W_n(\iota) = W$ are constant matrices being generators of continuous Markov chains for all n and all $\iota \in \mathcal{M}$, and $\widehat{W}_n \equiv 0$, (4.7) reduces to the existing standard consensus algorithm with additive noise. The nonadditive portion is a general nonlinear function of the analog state x, the Markov

chain state $\iota \in \mathcal{M}$, the noise source ζ, as well as n.

The noise sequences are "exogenous" in that (loosely speaking) the distribution of their future evolution, conditioned on their past, do not change if we also condition on the past of the state values. The distribution of the computation interval Y_n^i is allowed to depend on the state $x_{\tau_n^i}$ and the noise $\widetilde{\xi}_n^i$ which is used during that $(n+1)$-st interval in the ith processor. We define

$$\begin{aligned} &N_i(n) = \sup\{j : \tau_j^i \le n\}, \quad \Delta_n^i = n - \tau^i_{N_i(n)}, \\ &\xi_n^i = \widetilde{\xi}_j^i \quad \zeta_n^i = \widetilde{\zeta}_j^i, \quad \text{for} \quad n \in [\tau_j^i, \tau_{j+1}^i), \\ &\widetilde{x}_n = (\widetilde{x}_n^1, \ldots, \widetilde{x}_n^r) \quad \text{where} \quad \widetilde{x}_n^i \stackrel{\text{def}}{=} (x^1_{\tau^i_{N_i(n)}}, \ldots, x^r_{\tau^i_{N_i(n)}})' \\ &\text{and with a slight abuse of notation, denote} \quad \alpha_n = \widetilde{\alpha}_{\tau^i_{N_i(n)}}. \end{aligned} \tag{4.8}$$

Note that $\Delta_n^i = 0$ if n is a renewal time for processor i (Δ_n^i is the time elapsed since the start of a new computation for processor i). The $\widetilde{x}_n$ is an aggregate vector of dimension $r \cdot r$ and $\widetilde{x}_n^i$ is the state value used for the ith processor at real time n. We can now write

$$x_{n+1}^i = x_n^i + \mu[M_n(\alpha_n)\widetilde{x}_n^i]^i I_{n+1}^i + \mu[W_n(\alpha_n)\xi_n^i]^i I_{n+1}^i + \mu\widehat{W}_n^i(\widetilde{x}_n^i, \alpha_n, \zeta_n^i) I_{n+1}^i, \tag{4.9}$$

where $I_n^i = I_{\{\Delta_n^i = 0\}}$. If $I_n^i = 1$, then n is a random computation time for the ith processor. Note that the dependence of $\widetilde{\alpha}_{\tau^i_{N_i(n)}}$ is only through the random computation time τ_j^i with $j = N_i(n)$. Thus in the notation of α_n, we suppressed the i dependence for notational simplicity in the subsequent calculation; this information is also reflected from I_n^i. In what follows, for each i, the mixing measures defined in (A2), namely, ϕ and ψ should be i dependent. Nevertheless, to simplify the notation, instead of writing ϕ^i and ψ^i, we will use ϕ and ψ throughout the rest of the chapter. We assume the following conditions hold.

(A1) The process $\widetilde{\alpha}_n$ is a discrete-time Markov chain with a finite state space $\mathcal{M} = \{1, \ldots, m_0\}$

and transition probability matrix

$$P = P^{\varepsilon} = I + \varepsilon Q, \tag{4.10}$$

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 $i = 1, \ldots, m_0$) such that Q is irreducible.

(A2) (a) (i) The function $\widehat{W}_n(\cdot, \iota, e)$ is continuous for each $\iota \in \mathcal{M}$, $e \in \mathbb{R}^r$, and each n, and $|\widehat{W}_n(x, \iota, e)| \leq K(1 + |x|)$ for each $x \in \mathbb{R}^r$, $\iota \in \mathcal{M}$, e, and n. (ii) The $\{\widetilde{\xi}_n^i\}$ is a sequence of $\mathbb{R}^r$-valued ϕ-mixing processes such that $E\widetilde{\xi}_n^i = 0$, $E|\widetilde{\xi}_n^i|^{2+\delta} < \infty$ for some $\delta > 0$. Denote $\mathcal{F}_n^{\widetilde{\xi}^i} = \sigma\{\widetilde{\xi}_k^i; k \leq n\}$, $\mathcal{F}^{\widetilde{\xi}^i, n} = \sigma\{\widetilde{\xi}_k^i;\ k \geq n\}$. For $m > 0$, the mixing measure is defined by $\phi(m) = \sup_{B \in \mathcal{F}^{\widetilde{\xi}^i, n+m}} |P(B|\mathcal{F}_n^{\widetilde{\xi}^i}) - P(B)|_{\frac{2+\delta}{1+\delta}}$ and it satisfies $\sum_{m=1}^{\infty} \phi^{\frac{\delta}{1+\delta}}(m) < \infty$. (iii) The $\{\widetilde{\zeta}_n^i\}$ is a stationary sequence that is uniformly bounded such that for each $x \in \mathbb{R}^r$ and each $\iota \in \mathcal{M}$ and each n, $E\widehat{W}_n(x, \iota, \widetilde{\zeta}_n^i) = 0$. Moreover, $\widehat{W}_n(x, \iota, \widetilde{\zeta}_n^i)$ is uniformly mixing such that the mixing measure satisfies $\sum_m \psi^{1/2}(m) < \infty$. (iv) The sequences $\{\widetilde{\alpha}_n\}$, $\{\widetilde{\xi}_n^i\}$, and $\{\widetilde{\zeta}_n^i\}$ are independent.

(b) For each i, the sequence of positive integer-valued random variables $\{Y_n^i\}$ is bounded. There are $\widetilde{\pi}^i(x, \iota, \widetilde{\xi}^i)$ continuous in x uniformly in each bounded $(x, \widetilde{\xi}^i)$ set such that$\widetilde{E}_{\tau_n^i} Y_n^i = \widetilde{\pi}^i(x_{\tau_n^i}, \iota, \widetilde{\xi}_n^i)$,where$\widetilde{E}_m^i$denotes the conditional expectation on $\widetilde{\mathcal{F}}_m^i = \{x_0, \widetilde{\alpha}_{\tau_j^i}, \widetilde{\xi}_{j-1}^i, \widetilde{\zeta}_{j-1}^i : j \leq m\}$. There are continuous and strictly positive $\pi^i(\cdot, \iota)$ such that for each x and each m, $\sum_{j=m}^{n+m} E_m^i \widetilde{\pi}^i(x, \iota, \widetilde{\xi}_j^i)/n \to \pi^i(x, \iota)$ in probability, for each x and $\iota \in \mathcal{M}$.

(c) For each $\iota \in \mathcal{M}$, $\{M_n(\iota)\}$ and $\{W_n(\iota)\}$ are uniformly bounded. For each $\iota \in \mathcal{M}$,

there is an $\overline{M}(\iota)$ such that for each m, $\sum_{j=m}^{m+n-1} M_j(\iota)/n \to \overline{M}(\iota)$, where $\overline{M}(\iota)$ is an irreducible generator of a Markov chain for each $\iota \in \mathcal{M}$.

Remark 4.2. (a) Note that as a consequence of (A2), for any positive integer m and fixed ι,

$$\begin{aligned} &\frac{1}{n}\sum_{j=m}^{m+n-1} \widetilde{E}^i_m \widehat{W}_j(x,\iota,\widetilde{\zeta}^i_j) \to 0 \text{ in probability,} \\ &\frac{1}{n}\sum_{j=m}^{m+n-1} \widetilde{E}^i_m \widetilde{\xi}^i_j \to 0 \text{ in probability.} \end{aligned} \tag{4.11}$$

In what follows, we often work with ξ^i_n and ζ^i_n. Then we use E^i_m to denote the conditioning on the σ-algebra $\mathcal{F}^i_m = \{x_0, \alpha_j, \xi^i_{j-1}, \zeta^i_{j-1} : j \le m\}$.

(b) Assuming that $\pi^i(\cdot,\iota)$ is strictly positive is reasonable. This is essentially a suitably scaled limit of the mean of Y^i_n. Under the standard renewal setup with i.i.d. inter-arrival Y^i_n (independent of data), it is simply a positive constant, the mean of Y^i_n.

(c) Note that (4.9) is a stochastic approximation type algorithm, but more difficult to analyze because of the switching topologies. In the traditional setup of stochastic approximation problems, the limit or the averaged system is an ordinary differential equation (ODE). Very often these limits are autonomous. Even if they are time inhomogeneous ODE, these equations are non-random. Certain cases treated here, the limits is no longer an ODE, but a randomly varying ODE subject to switching. 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 [134]).

In the next two sections, three possibilities concerning the relative sizes of ε and μ are analyzed. This idea also appears in related treatments of LMS-type algorithms under regime-switching dynamic systems, see [125–127].

In treating the three different cases, careful analysis is needed to examine convergence, stability, and related consensus issues.

4.4 Asymptotic Properties: $\varepsilon = O(\mu)$

This section concentrates on the case $\varepsilon = O(\mu)$. For notational simplicity and concreteness, in what follows, we simply consider $\varepsilon = \mu$ in this section. More general cases can be considered; they do not add further technical difficulties. The results will be similar in spirit.

4.4.1 Basic Properties

To proceed, we first present a moment estimate for the recursive algorithm (4.9). Throughout the chapter, we use K to denote a generic positive constant with the convention $K+K = K$ and $KK = K$. We also use K_T to denote a generic positive constant that depends on T (whose value may change for different appearances).

Lemma 4.3. *Under Assumption* (A1)*, for any* $0 < T < \infty$ *and each* $i = 1, \ldots, r$*,*

$$\sup_{0 \le n \le T/\varepsilon} E|x_n^i|^2 \le K_T \exp(T) < \infty.$$

Proof. Note that for any $0 < T < \infty$ and $0 \le n \le T/\mu$, by Cauchy-Schwarz inequality,

$$\begin{aligned} \mu^2 E \left| \sum_{k=0}^{n} [M_k(\alpha_k)\widetilde{x}_k^i]^i I_{k+1}^i \right|^2 &\le \mu^2 (n+1) E \sum_{k=0}^{n} |M_k(\alpha_k)|^2 \left| [\widetilde{x}_k^i]^i I_{k+1}^i \right|^2 \\ &\le K_T \mu \sum_{k=0}^{n} E \left| [\widetilde{x}_k^i]^i I_{k+1}^i \right|^2, \end{aligned} \tag{4.12}$$

where $K_T > 0$. Likewise,

$$\begin{aligned}
\mu^2 E\left|\sum_{k=0}^{n} \widehat{W}_k(\widetilde{x}_k^i, \alpha_k, \zeta_k^i)\right|^2 &\le K_T \mu \sum_{k=0}^{n} E\left|[\widetilde{x}_k^i]^i I_{k+1}^i\right|^2 + K_T \\
\mu^2 E\left|\sum_{k=0}^{n} [W_k(\alpha_k)\xi_k^i]^i I_{k+1}^i\right|^2 &\le K(\mu n)^2 \le K_T.
\end{aligned} \tag{4.13}$$

Iterating on $E|x_n^i|^2$ with the use of (4.9) and using (4.12) and (4.13), we obtain

$$\begin{aligned}
E|x_{n+1}^i|^2 &\le (E|x_0^i|^2 + K_T) + K_T \mu \sum_{k=0}^{n} E\left|[\widetilde{x}_k^i]^i I_{k+1}^i\right|^2 \\
&\le (E|x_0^i|^2 + K_T) + K_T \mu \sum_{k=0}^{n} E|x_k^i|^2 + O(\mu).
\end{aligned} \tag{4.14}$$

Then by Gronwall's inequality,

$$E|x_{n+1}^i|^2 \le K_T \exp(n\mu) \le K_T \exp(\mu(T/\mu)) \le K_T \exp(T).$$

Taking sup over n, the desired estimate follows. □

4.4.2 Convergence

This section is devoted to obtaining asymptotic properties of algorithm (4.9). Before proceeding further, we state a result on estimation error bounds. The proof of the assertion on probability distributions is essentially in that of Theorem 3.5 and Theorem 4.3 of [131], whereas the proof of weak convergence of $\alpha^\varepsilon(\cdot)$ can be found in [133]; see also [132]. Thus the proof is omitted.

Lemma 4.4. *Under condition* (A2)*, with* P^ε *given by* (4.10)*, denote the* n*-step transition probability by* $(P^\varepsilon)^n$ *and* $p_n^\varepsilon = (P(\widetilde{\alpha}_n = 1), \ldots, P(\widetilde{\alpha}_n = m_0))$*, and define* $\alpha^\varepsilon(t) = \widetilde{\alpha}_n$ *for*

$t \in [n\varepsilon,\ n\varepsilon + \varepsilon)$. Then the following claims hold:

$$\begin{aligned} p_n^\varepsilon &= p(t) + O(\varepsilon + e^{-k_0 t/\varepsilon}), \\ (P^\varepsilon)^{n-n_0} &= \Xi(\varepsilon n, \varepsilon n_0) + \mathcal{O}(\varepsilon + e^{-k_0(n-n_0)}), \end{aligned} \tag{4.15}$$

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

$$\begin{aligned} \frac{dp(t)}{dt} &= p(t)Q,\ P(0) = p_0, \\ \frac{\Xi(t, t_0)}{dt} &= \Xi(t, t_0)Q,\ \Xi(t_0, t_0) = I, \end{aligned} \tag{4.16}$$

with $t_0 = \varepsilon\, n_0$ and $t = \varepsilon\, n$. Moreover, $\alpha^\varepsilon(\cdot)$ converges weakly to $\alpha(\cdot)$, a continuous-time Markov chain generated by Q.

Since we consider $\varepsilon = O(\mu)$, without loss of generality, we take $\varepsilon = \mu$ in what follows. The next lemma concerns the property of the algorithm as $\mu \to 0$ through an appropriate continuous-time interpolation. We define

$$x^\mu(t) = x_n,\ \alpha^\mu(t) = \widetilde{\alpha}_n, \quad \text{for} \quad t \in [\mu n, \mu n + \mu).$$

Then $(x^\mu(\cdot), \alpha^\mu(\cdot)) \in D([0, T] : \mathbb{R}^r \times \mathcal{M})$, which is the space of functions that are defined on $[0, T]$ taking values in $\mathbb{R}^r \times \mathcal{M}$, and that are right continuous and have left limits endowed with the Skorohod topology [109, Chapter7]. Before proceeding further, we first state a lemma that gives the uniqueness of the solution of (4.17).

Lemma 4.5. *The switched ordinary differential equation*

$$\frac{dx^i(t)}{dt} = \frac{[\overline{M}(\alpha(t))x(t)]^i}{\pi^i(x(t), \alpha(t))},\ i = 1, \ldots, r. \tag{4.17}$$

has a unique solution for each initial condition $(x(0), \alpha(0))$ *with* $x(0) = (x_0^1, \ldots, x_0^r)'$.

Proof. For any $f(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}$ satisfying for each $\iota \in \mathcal{M}$, $f(\cdot, \iota) \in C_0^1$ (space of continuously differentiable functions with compact support), $\mathcal{L}_1$ is defined as follows:

$$\mathcal{L}_1 f(x,\iota) = \sum_{i=1}^{r} \frac{\partial f(x,\iota)}{\partial x_i} \frac{[\overline{M}(\iota)x]^i}{\pi^i(x,\iota)} + Qf(x,\cdot)(\iota), \quad \iota \in \mathcal{M}, \tag{4.18}$$

where

$$Qf(x,\cdot)(\iota) = \sum_{\ell=1}^{m_0} q_{\iota\ell} f(x,\ell).$$

Let $(x(t), \alpha(t))$ be a solution of the martingale problem with operator $\mathcal{L}_1$ defined in (4.18). We proceed to show that the solution is unique in the sense of 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,

$$\psi_{j_0k_0}(t) - \psi_{j_0k_0}(0) - \int_0^t \mathcal{L}_1 \psi_{j_0k_0}(s) ds = 0, \tag{4.19}$$

where

$$\mathcal{L}_1 \psi_{j_0k_0}(s) = \sum_{i=1}^{m_0} \gamma_i \frac{[\overline{M}(k_0)x]^i}{\pi^i(x(s),\alpha(s))} \psi_{j_0k_0}(s) + \sum_{j=0}^{m_0} q_{jj_0} \psi_{jk_0}(s). \tag{4.20}$$

Let $\psi(t) = (\psi_{\iota\ell}(t) : \iota \le m_0, \ell \le m_0)$. Combining (4.19) and (4.20), we obtain

$$\psi(t) = \psi(0) + \int_0^t G\psi(s) ds, \tag{4.21}$$

where G is an $m_0 \times m_0$ matrix. Thus (4.21) is an ordinary differential equation with an initial condition $\psi(0)$. As a result, it has a unique solution. □

By Lemma 4.3 to 4.5, we can obtain the following theorem.

Theorem 4.6. *Under* (A1) *and* (A2), $\{x^\mu(\cdot), \alpha^\mu(\cdot)\}$ *is tight in* $D([0,T] : \mathbb{R}^r \times \mathcal{M})$. *Assume that* x_0 *and* α_0 *to be independent of* μ *and are non-random without loss of generality. Then* $(x^\mu(\cdot), \alpha^\mu(\cdot))$ *converges weakly to* $(x(\cdot), \alpha(\cdot))$ *that is a solution of* (4.17) *with initial condition* (x_0, α_0).

Proof. (a) Tightness. The tightness of $\{\alpha^\mu(\cdot)\}$ can be proved as in that of [131, Theorem 4.3]. So we need only prove the tightness of $\{x^\mu(\cdot)\}$, i.e., we need only prove that $\{x^{\mu,i}(\cdot)\}$ is tight for each i.

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

$$x^{\mu,i}(t+s) - x^{\mu,i}(t) = \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} [M_k(\alpha_k)\widetilde{x}_k^i]^i I_{k+1}^i + \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} [W_k(\alpha_k)\xi_k^i]^i I_{k+1}^i + \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} [\widehat{W}_k^i(\widetilde{x}_k^i, \alpha_k, \zeta_k^i)] I_{k+1}^i.$$

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

Since $\widetilde{\alpha}_k$ is a finite-state Markov chain, by (A1) $|M_k(\alpha_k)|$ and $|W_k(\alpha_k)|$ (see the notation α_k in (4.8)) are uniformly bounded. Using the Cauchy-Schwarz inequality as in (4.12) (with

$\sum_{k=0}^{n}$ replaced by $\sum_{k=t/\mu}^{(t+s)/\mu-1}$) together with Lemma 4.3,

$$
\begin{aligned}
E_t^\mu \left|x^{\mu,i}(t+s)-x^{\mu,i}(t)\right|^2 \le K\mu^2 E_t^\mu \left[\left|\sum_{k=t/\mu}^{(t+s)/\mu-1}[M_k(\alpha_k)\widetilde{x}_k^i]^i I_{k+1}^i\right|^2+\left|\sum_{k=t/\mu}^{(t+s)/\mu-1}[W_k(\alpha_k)\xi_k^i]^i I_{k+1}^i\right|^2\right] \\
+K\mu^2 E_t^\mu \left|\sum_{k=t/\mu}^{(t+s)/\mu-1}[\widehat{W}(\widetilde{x}_k^i,\alpha_k,\zeta_k^i)]^i I_{k+1}^i\right|^2 \\
\le K\mu s \sum_{k=t/\mu}^{(t+s)/\mu-1}\sup_{t/\mu\le k\le (t+s)/\mu-1} E_t^\mu|x_k^i|^2+Ks^2\le K\delta^2,
\end{aligned}
\tag{4.22}
$$

where E_t^μ denotes the conditioning on $\mathcal{F}_t^\mu=\sigma\{x_0^\mu,\xi_k^i,\zeta_k^i: i\le r, k<\lfloor t/\mu\rfloor\}$. In the above, we have used $E_t^\mu|x_k|^2<\infty$ for $\lfloor t/\mu\rfloor\le k<\lfloor (t+s)/\mu\rfloor$, which can be shown as in Lemma 4.3. As a result,

$$
\lim_{\delta\to 0}\limsup_{\mu\to 0} E\left[\sup_{0\le s\le\delta} E_t^\mu|x^{\mu,i}(t+s)-x^{\mu,i}(t)|^2\right]=0.
$$

The tightness of $\{x^{\mu,i}(\cdot)\}$ follows from [106, p.47].

(b) Characterization the limit. For notational simplicity, we shall not use a function $f(\cdot,\cdot)\in C_0^2$ in the usual martingale problem formulation for the following derivation, but work with the underlying sequences directly. It is convenient to proceed with a scaling argument to treat the random renewal times.

Define the process $Z_n^{\mu,i}$, $Z^{\mu,i}(\cdot)$, $\Psi_n^{\mu,i}$, and $\Psi^{\mu,i}(\cdot)$ by

$$
\begin{aligned}
&Z_n^{\mu,i}=\mu\sum_{j=0}^{n-1}Y_j^i,\quad Z^{\mu,i}(t)=Z_n^{\mu,i}\text{ on }[n\mu,\,n\mu+\mu),\\
&\Psi_n^{\mu,i}=\mu\sum_{k=0}^{n-1}[M_{\tau_k^i}(\widetilde{\alpha}_{\tau_k^i})x_{\tau_k^i}]^i+\mu\sum_{k=0}^{n-1}W_{\tau_k^i}(\widetilde{\alpha}_{\tau_k^i})\widetilde{\xi}_k^i+\mu\sum_{k=0}^{n-1}\widehat{W}_{\tau_k^i}^i(x_{\tau_k^i}^i,\widetilde{\alpha}_{\tau_k^i},\widetilde{\zeta}_k^i),\\
&\Psi^{\mu,i}(t)=\Psi_n^{\mu,i}\ \text{ for }\ t\in[n\mu,n\mu+\mu),
\end{aligned}
$$

Use the method similar as (a), we can prove that $Z^{\mu,i}(\cdot)$ and $\Psi^{\mu,i}(\cdot)$ are tight. As a result,

we obtain that $\{x^{\mu,i}(\cdot), Z^{\mu,i}(\cdot), \Psi^{\mu,i}(\cdot)\}$ is tight in $D([0,\infty): \mathbb{R}^3)$ and all limits are uniformly Lipschitz continuous. We fix and work with a weakly convergent subsequence, also indexed by μ, and with limit denoted by $(x^i(\cdot), Z^i(\cdot), \Psi^i(\cdot))$, for $i \leq r$. Now we state a lemma.

Lemma 4.7. *Under the conditions of Theorem* 4.6*, the limits of* $Z^{\mu,i}(\cdot)$ *and* $\Psi^{\mu,i}(\cdot)$ *satisfy*

$$Z^i(t) = \int_0^t \pi^i(x(Z^i(s)), \alpha(Z^i(s)))ds, \tag{4.23}$$

and

$$\Psi^i(t) = \int_0^t [\overline{M}(\alpha(Z^i(u))x(Z^i(u))]^i du, \tag{4.24}$$

respectively.

Proof of Lemma 4.7. Fixed i, using similar argument as that of [108, pp. 224-226], we can derive (4.23). The details are omitted.

Next we work on $\Psi^{\mu,i}(\cdot)$ and concentrate on the proof of (4.24). First, it is readily seen that for any $t, s > 0$,

$$\Psi^{\mu,i}(t+s) - \Psi^{\mu,i}(t) = \mu \sum_{\ell=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu} \left\{ [M_{\tau_k^i}(\ell) x_{\tau_k^i}]^i + [W_{\tau_k^i}(\ell)\widetilde{\xi}_k^i]^i + \widehat{W}_{\tau_k^i}^i(x_{\tau_k^i}^i, \ell, \widetilde{\zeta}_k^i) \right\} I_{\{\widetilde{\alpha}_{\tau_k^i} = \ell\}}. \tag{4.25}$$

Next, pick out any bounded and continuous function $h(\cdot)$, any positive integer κ, and any $t_j \leq t$ for $j \leq \kappa$, the weak convergence and Skorohod representation imply that as $\mu \to 0$,

$$\begin{aligned} &Eh(x^\mu(t_j), \alpha^\mu(t_j) : j \leq \kappa)[\Psi^{\mu,i}(t+s) - \Psi^{\mu,i}(t)] \\ &\quad \to Eh(x(t_j), \alpha(t_j) : j \leq \kappa)[\Psi^i(t+s) - \Psi^i(t)]. \end{aligned} \tag{4.26}$$

Choose m_μ so that $m_\mu \to \infty$ as $\mu \to 0$, but $\mu\, m_\mu = \delta_\mu \to 0$. By the continuity of

linear function in the variable x,

$$\begin{aligned}
&\lim_{\mu} Eh(x^{\mu}(t_j), \alpha^{\mu}(t_j) : j \le \kappa)\Big[\mu \sum_{\ell=1}^{m_0} \sum_{k=t/\mu}^{(t+s)/\mu} [M_{\tau_k^i}(\ell) x_{\tau_k^i}]^i I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}\Big] \\
&= \lim_{\mu} Eh(x^{\mu}(t_j), \alpha^{\mu}(t_j) : j \le \kappa)\Big[\sum_{\ell=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [M_{\tau_k^i}(\ell) x_{\tau_k^i}]^i I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}\Big] \\
&= \lim_{\mu} Eh(x^{\mu}(t_j), \alpha^{\mu}(t_j) : j \le \kappa)\Big[\sum_{\ell=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [M_{\tau_k^i}(\ell) x_{\tau_{lm_\mu}^i}]^i I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}\Big] \\
&= \lim_{\mu} Eh(x^{\mu}(t_j), \alpha^{\mu}(t_j) : j \le \kappa)\Big[\sum_{\ell=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [\overline{M}(\ell) x_{\tau_{lm_\mu}^i}]^i I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}\Big] \\
&\quad + \lim_{\mu} Eh(x^{\mu}(t_j), \alpha^{\mu}(t_j) : j \le \kappa)\Big[\sum_{\ell=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [(M_{\tau_k^i}(\ell) - \overline{M}(\ell)) x_{\tau_{lm_\mu}^i}]^i I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}\Big].
\end{aligned} \tag{4.27}$$

Denoting $P(\widetilde{\alpha}_{\tau_k^i} = \ell |\ \widetilde{\alpha}_{\tau_{lm_\mu}^i}) = p(\tau_k^i, \tau_{lm_\mu}^i)$ with ℓ suppressed, inserting conditional expectation and using a partial summation, we obtain that

$$\begin{aligned}
&\frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} (M_{\tau_k^i}(\ell) - \overline{M}(\ell)) p(\tau_k^i, \tau_{lm_\mu}^i) \\
&= \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} (M_{\tau_k^i}(\ell) - \overline{M}(\ell)) p(\tau_{lm_\mu+m_\mu-1}^i, \tau_{lm_\mu}^i) \\
&\quad + \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-2} \sum_{k=lm_\mu}^{j} (M_{\tau_k^i}(\ell) - \overline{M}(\ell)) [p(\tau_j^i, \tau_{lm_\mu}^i) - p(\tau_{j+1}^i, \tau_{lm_\mu}^i)].
\end{aligned} \tag{4.28}$$

By virtue of the assumption (A2)(c), the term on the second line of (4.28) goes to 0 as $\mu \to 0$. Next, noting

$$(I + \mu Q)^{\tau_{j+1}^i - \tau_{lm_\mu}^i} - (I + \mu Q)^{\tau_j^i - \tau_{lm_\mu}^i} = O(\mu),$$

we have

$$[p(\tau_j^i, \tau_{lm_\mu}^i) - p(\tau_{j+1}^i, \tau_{lm_\mu}^i)] \ = \ O(\mu).$$

As a result,

$$
\begin{aligned}
&E\left|h(x^\mu(t_j),\alpha^\mu(t_j):j\le\kappa)E_{\tau^i_{lm_\mu}}\Big[\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}[(M_{\tau^i_k}(\ell)-\overline{M}(\ell))x_{\tau^i_{lm_\mu}}]^iI_{\{\widetilde{\alpha}_{\tau^i_k}=\ell\}}\Big]\right|\\
&\le E\left|\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}[(M_{\tau^i_k}(\ell)-\overline{M}(\ell))x_{\tau^i_{lm_\mu}}]^iI_{\{\widetilde{\alpha}_{\tau^i_k}=\ell\}}\right|\\
&\le\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu\left|\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}(M_{\tau^i_k}(\ell)-\overline{M}(\ell))\right|E|x_{\tau^i_{lm_\mu}}|O(\mu)\\
&\to 0\ \text{ as }\mu\to 0.
\end{aligned}
$$

Since $\mu\, m_\mu\to 0$ as $\mu\to 0$, when $\mu\, lm_\mu\to u$, for all $lm_\mu\le k\le lm_\mu+m_\mu$, $\mu\, k\to u$ as well. Therefore, the detailed estimates above lead to

$$
\begin{aligned}
&\lim_\mu Eh(x^\mu(t_j),\alpha^\mu(t_j):j\le\kappa)\Big[\mu\sum_{\ell=1}^{m_0}\sum_{k=t/\mu}^{(t+s)/\mu}[M_{\tau^i_k}(\ell)x_{\tau^i_{lm_\mu}}]^iI_{\{\widetilde{\alpha}_{\tau^i_k}=\ell\}}\Big]\\
&=\lim_\mu Eh(x^\mu(t_j),\alpha^\mu(t_j):j\le\kappa)\Big[\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu[\overline{M}(\ell)x_{\tau^i_{lm_\mu}}]^i\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}I_{\{\widetilde{\alpha}_{\tau^i_k}=\ell\}}\Big]\\
&=\lim_\mu Eh(x^\mu(t_j),\alpha^\mu(t_j):j\le\kappa)\Big[\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}[\overline{M}(\widetilde{\alpha}_{\tau^i_k})x_{\tau^i_{lm_\mu}}]^i\Big]\\
&=\lim_\mu Eh(x^\mu(t_j),\alpha^\mu(t_j):j\le\kappa)\Big[\sum_{\ell=1}^{m_0}\sum_{l\delta_\mu=t}^{t+s}\delta_\mu\\
&\qquad\qquad\times\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}[\overline{M}(\alpha^\mu(Z^{\mu,i}(\mu lm_\mu)))x^\mu(Z^{\mu,i}(\mu lm_\mu))]^i\Big]\\
&=Eh(x(t_j),\alpha(t_j):j\le\kappa)\Big[\int_t^{t+s}[M(\alpha(Z^i(u)))x(Z^i(u))]^idu\Big].
\end{aligned}
\tag{4.29}
$$

Next using the independence of $\widetilde{\alpha}_n$ with $\widetilde{\xi}^i_n$, similar conditioning argument together with the mixing conditions (see the mixing inequality given in [100, Corollary 2.4]) given in (A2)

(a) yields

$$\mu \sum_{k=t/\mu}^{(t+s)/\mu} E|W_{\tau_k^i}(\ell)\widetilde{E}^i_{t/\mu}\widetilde{\xi}^i_k|P(\widetilde{\alpha}_k=\ell|\widetilde{\alpha}_{t/\mu}) \le K\mu \sum_{k=t/\mu}^{(t+s)/\mu} \phi^{\delta/(1+\delta)}(k-(t/\mu)) \to 0 \ \text{ as } \ \mu \to 0. \tag{4.30}$$

Likewise, using the continuity of $\widehat{W}\ (\cdot, \ell, \widetilde{\zeta}^i)$, the limit of

$$\sum_{\ell=1}^{m_0} \mu \sum_{k=t/\mu}^{(t+s)/\mu} \widehat{W}_{\tau_k^i}(\ x^i_{\tau_k^i}, \ell, \ \widetilde{\zeta}^i_k) I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}$$

is the same as that of

$$\sum_{\ell=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \widehat{W}_{\tau_k^i}(x^i_{\tau^i_{lm_\mu}}, \ell, \widetilde{\zeta}^i_k) I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}}.$$

Then using the uniform mixing (see [97, p. 166]) given in (A2) to the last expression, we obtain

$$\mu \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} E|E_{\tau^i_{lm_\mu}} \widehat{W}_{\tau_k^i}(x^i_{\tau^i_{lm_\mu}}, \ell, \widetilde{\zeta}^i_k)|P(\widetilde{\alpha}_{\tau_k^i}=\ell|\widetilde{\alpha}_{t/\mu}) \to 0 \ \text{ as } \ \mu \to 0.$$

Thus

$$Eh(x^\mu(t_j), \alpha^\mu(t_j) : j \le \kappa) \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \widehat{W}_{\tau_k^i}(x^i_{\tau^i_{lm_\mu}}, \ell, \widetilde{\zeta}^i_k) I_{\{\widetilde{\alpha}_{\tau_k^i}=\ell\}} \to 0. \tag{4.31}$$

Using the estimates obtained thus far together with (4.25), we have proved that

$$Eh(x(t_j), \alpha(t_j) : j \le \kappa)\Big[\Psi^i(t+s) - \Psi^i(t) - \int_t^{t+s} [\overline{M}(\alpha(Z^i(u)))x(Z^i(u))]^i du\Big] = 0. \tag{4.32}$$

Therefore, the proof of the lemma is concluded. □

Completion of Proof of Theorem 4.6. With Z^{-1} denoting the inverse of Z, we have

$$\begin{aligned} x^{\mu,i}(t) - x^i(0) &= \mu \sum_{k=0}^{N_i(t/\mu)-1} \left\{ [M_k(\ell)\widetilde{x}_k^i]^i + [W_k(\ell)\xi_k^i]^i + \widehat{W}_k(\widetilde{x}_k^i, \ell, \zeta_k^i) \right\} I_{k+1}^i I_{\{\alpha_k=\ell\}} \\ &= \mu \sum_{\mu k=0}^{(Z^{\mu,i})^{-1}(t)} \left\{ [M_k(\ell)\widetilde{x}_k]^i + [W_k(\ell)\xi_k^i]^i + \widehat{W}_k(\widetilde{x}_k^i, \ell, \zeta_k^i) \right\} I_{k+1}^i I_{\{\alpha_k=\ell\}} \end{aligned} \tag{4.33}$$

Lemma 4.7 then yields the limit process

$$x^i(t) = x^i(0) + \int_0^{(Z^i)^{-1}(t)} [\overline{M}(\alpha(Z^i(u)))x(Z^i(u))]^i du.$$

This in turn implies

$$\begin{aligned} \dot{x}^i(t) &= [\overline{M}(\alpha(Z^i((Z^i)^{-1}(t))))(x(Z^i((Z^i)^{-1}(t)))]^i(\dot{Z}^i)^{-1}(t) \\ &= \frac{[\overline{M}(\alpha(t))x(t)]^i}{\pi^i(x(t), \alpha(t))} \end{aligned}$$

as desired. Thus the theorem is proved. □

4.5 Invariance Theorem

To study the long-time behavior, we derive an invariant theorem for the switched system. Following the discussion in [134, Chapter 9], recall that a Borel measurable set $U \subset \mathbb{R}^r \times \mathcal{M}$ is invariant with respect to the process $(x(t), \alpha(t))$ if $P((x(t), \alpha(t)) \in U, \text{ for all } t \geq 0) = 1$, for any initial $(x, \iota) \in U$. That is, a process starting from U will remain in U with probability 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, ι), $P(\sup_{t\geq 0} d(x(t), K_c) < \rho) \geq 1-\delta$ whenever $d(x, K_c) < \delta_1$;

- A closed and bounded set $K_c \subset \mathbb{R}^r$ is said to be asymptotically stable in probability if it is stable in probability, and $P(\lim_{t\to\infty} d(x(t), K_c) = 1) \to 1$, 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.

Theorem 4.8. *Assume that for each $\iota \in \mathcal{M}$, $\overline{M}(\iota)$ is irreducible. Under the conditions of Theorem* 4.6*, 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.*

Proof. To prove (i), we divide the time intervals according to the associated switch times. We begin with $(x(0), \alpha(0)) = (x_0, \iota)$. Following the dynamic system given in (4.17), let ρ_1 be the first switching time, i.e., $\rho_1 = \inf\{t : \alpha(t) = \iota_1 \neq \iota\}$. Note that $x(t) = x(t, \omega)$, where $\omega \in \Omega$ is the sample point. Then in the interval $[0, \rho_1)$, for almost all ω, (4.17) is a system with constant matrix $\overline{M}(\iota)$. For all $t \in [0, \rho_1]$, from equation (4.17) we have that for any $T < \infty$ and $t \in [0, T]$,

$$x(t) \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{d^k\, x\,(0)}{dt^k}, \qquad \text{and} \qquad \sup_{0\leq\, t\leq\, T} \left| \frac{d^k\, x\,(0)}{dt^k} \right| \leq\; K \;< \infty.$$

Because the matrix $\overline{M}(\iota)$ is irreducible, there is an eigenvalue 0 and the rest of the eigenvalues all have negative real parts.

If $x(0) \in \mathsf{Z}$, then $x(0) = c\mathbb{1}$, $[\overline{M}(\iota)x(0)]^i = 0$, and for each $i \leq r$,

$$\frac{dx^i}{dt}(0) = \frac{[\overline{M}(\iota)x(0)]^i}{\pi^i(x(0), \iota)} = 0,$$

and similarly, we obtain

$$\frac{d^k x^i}{dt^k}(0) = 0 \quad \text{for all} \quad k > 0\,, \quad i \leq r.$$

Therefore, $x(t) = x(0)$ for all $t \in [0, \rho_1)$. Thus $x(t) \in \mathsf{Z}$ for all $t \in [0, \rho_1)$. Now, define $\rho_2 = \inf\{t \geq \rho_1 : \alpha(t) = \iota_2 \neq \iota_1\}$. By the continuity of $x(\cdot)$, $x(\rho_1) = x(\rho_1^-) \in \mathsf{Z}$. Similar as in the previous paragraph, we can show for all $t \in [\rho_1, \rho_2)$, $x(t) \in \mathsf{Z}$. Continue in this way. For any $T > 0$, consider $[0, T]$. Then $0 < \rho_1 < \rho_2 < \cdots < \rho_{N(T)} \leq T$, where $N(t)$ is the counting process that counts the number of switchings in the interval $[0, T]$, and ρ_n is defined recursively such that $\alpha(\rho_n) = \iota_n$ and $\rho_{n+1} = \inf\{t \geq \rho_n : \alpha(t) = \iota_{n+1} \neq \iota_n\}$. Suppose that we have for all $t \leq \rho_n$, $x(t) \in \mathsf{Z}$ w.p.1. Using induction, we can show $x(t) \in \mathsf{Z}$ for all $t \in [0, \rho_{N(T)})$. Finally, we work with the interval $[\rho_{N(T)}, T]$, this establishes the first assertion.

To prove (ii), define $V(x) = x'x/2$. Since $V(x)$ is independent of the switching component, $\sum_{\ell=1}^{m_0} q_{\iota\ell} V(x) = 0$. Thus, for each $\iota \in \mathcal{M}$, because of the irreducibility of $\overline{M}(\iota)$,

$$\mathcal{L}_1 V(x) = \sum_{i=1}^{r} \frac{x^i [\overline{M}(\iota)x]^i}{\pi^i(x, \iota)} < 0, \quad \text{for all} \quad x \notin \mathsf{Z}.$$

The rest of the proof of the stability in probability of the set Z is similar in spirit to that of [134, Chapter 9]. We omit the details for brevity. □

Denote $x_c = \eta\, \mathbb{1}$ With the above proposition, we can further obtain the following result as a corollary of Theorem 4.8.

Corollary 4.9. *Assume the conditions of Theorem* 4.8. *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* $\delta > 0$, $\lim_{\mu \to 0} P(|x^\mu(\cdot + t_\mu) - x_c| \geq \delta) = 0$.

4.5.1 Normalized Error Sequences

This section is devoted to analyzing the rates of variations of scaled sequence of errors. We begin with a result on upper bounds on estimation errors in the mean square sense.

Theorem 4.10. *Assume the conditions of Theorem* 4.6. *Then there is an* N_μ *such that* $E\,|x_n|^2 = O(1)$ *for all* $n \geq N_\mu$.

Proof. We prove the assertion by means of perturbed Liapunov function methods. Redefine $V(x) = (x - x_c)'(x - x_c)/2$. Note that $V_{x^i}(x) = (\partial/\partial\, x^i)V(x) = (x^i - x_c^i)$ and $V_{xx}(x) = I$ the identity matrix. Using a Taylor expansion for $V(x)$, we have

$$
\begin{aligned}
&E_n V(x_{n+1}) - V(x_n) \\
&\quad = \mu \sum_{i=1}^{r} (x_n^i - x_c^i) \Big\{ (\overline{M}(\alpha_n)(x_n - x_c))^i + [(M(\alpha_n) - \overline{M}(\alpha_n))(x_n - x_c)]^i \\
&\qquad\qquad + (M_n(\alpha_n)(\widetilde{x}_n^i - x_n))^i + [W_n(\alpha_n)\xi_n^i]^i + \widehat{W}_n^i(\widetilde{x}_n^i, \alpha_n, \zeta_n^i) \\
&\qquad\qquad + [(M(\alpha_n) - \overline{M}(\alpha_n))x_c]^i \Big\} I_{n+1}^i + O(\mu^2)(V(x_n) + 1).
\end{aligned} \tag{4.34}
$$

Note that for all $x_n \not\in \mathsf{Z}$, $\sum_{\ell=1}^{m_0} x_n^i (\overline{M}(\ell)x_n)^i I_{n+1}^i I_{\{\alpha_n = \ell\}} \leq -\lambda V(x_n)$ for some $\lambda > 0$. For

some $0 < \lambda_0 < 1$, define

$$
\begin{aligned}
V_1^\mu(x,n) &= \mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n}^{\infty} \lambda_0^{j-n} E_n(x^i - x_c^i)[(M(\ell) - \overline{M}(\ell))(x - x_c)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i, \\
V_2^\mu(x,n) &= \mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n}^{\infty} \lambda_0^{j-n} E_n(x^i - x_c^i)[M_j(\ell)(\widetilde{x}^i - x)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i, \\
V_3^\mu(x,n) &= \mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n}^{\infty} E_n(x^i - x_c^i)[W_j(\ell)\xi_j^i)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i, \\
V_4^\mu(x,n) &= \mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n}^{\infty} E_n(x^i - x_c^i)\widehat{W}_j^i(\widetilde{x}^i, \ell, \zeta_j^i)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i, \\
V_5^\mu(x,n) &= \mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n}^{\infty} \lambda_0^{j-n} E_n(x^i - x_c^i)[(M(\ell) - \overline{M}(\ell))x_c]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i.
\end{aligned}
\tag{4.35}
$$

It is easily checked that

$$
|\, V_i^\mu(x,n)| \;=\; O(\mu)(\, V(x) + 1), \qquad i \;= 1, \ldots,\; 5. \tag{4.36}
$$

Noting

$$
\begin{aligned}
&\mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n+1}^{\infty} \lambda_0^{j-(n+1)} E_n(x_{n+1}^i - x_c^i)[(M(\ell) - \overline{M}(\ell))(x_{n+1} - x_c)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i \\
&-\mu \sum_{i=1}^{r} \sum_{\ell=1}^{m_0} \sum_{j=n+1}^{\infty} \lambda_0^{j-(n+1)} E_n(x_n^i - x_c)[(M(\ell) - \overline{M}(\ell))(x_n - x_c)]^i I_{\{\alpha_j=\ell\}} I_{j+1}^i \\
&\quad = O(\mu^2)(V(x_n) + 1),
\end{aligned}
$$

we have

$$
\begin{aligned}
&E_n V_1^\mu(x_{n+1}, n+1) - V_1^\mu(x_n, n) \\
&\quad = E_n[V_1^\mu(x_{n+1}, n+1) - V_1^\mu(x_n, n+1)] + [E_n V_1^\mu(x_{n+1}, n+1) - V_1^\mu(x_n, n)] \\
&\quad = -\mu \sum_{i=1}^{r} E_n(x_n^i - x_c^i)[(M(\alpha_n) - \overline{M}(\alpha_n)]x_n)^i I_{n+1}^i + O(\mu^2)(V(x_n) + 1).
\end{aligned}
\tag{4.37}
$$

Likewise, we obtain

$$
\begin{aligned}
&E_n V_2^\mu(x_{n+1}, n+1) - V_2^\mu(x_n, n) \\
&\quad = -\mu \sum_{i=1}^{r} E_n(x_n^i - x_c^i)[(M(\alpha_n)(\widetilde{x}_n^i - x_n)]^i I_{n+1}^i + O(\mu^2)(V(x_n)+1), \\
&E_n V_3^\mu(x_{n+1}, n+1) - V_3^\mu(x_n, n) \\
&\quad = -\mu \sum_{i=1}^{r} E_n(x_n^i - x_c^i)[W_n(\alpha_n)\xi_n^i]^i I_{n+1}^i + O(\mu^2)(V(x_n)+1), \\
&E_n V_4^\mu(x_{n+1}, n+1) - V_4^\mu(x_n, n) \\
&\quad = -\mu \sum_{i=1}^{r} E_n(x_n^i - x_c^i)\widehat{W}_n^i(\widetilde{x}_n, \alpha_n, \zeta_n^i) I_{n+1}^i + O(\mu^2)(V(x_n)+1), \\
&E_n V_5^\mu(x_{n+1}, n+1) - V_5^\mu(x_n, n) \\
&\quad = -\mu \sum_{i=1}^{r} E_n(x^i - x_c^i)[(M(\alpha_n) - \overline{M}(\alpha_n))x_c]^i I_{n+1}^i + O(\mu^2)(V(x_n)+1).
\end{aligned}
\tag{4.38}
$$

Define $V^\mu(x,n) = V(x) + \sum_{l=1}^{5} V_l^\mu(x,n)$. Using (4.34), (4.37), and (4.38), we arrive at

$$
E_n\, V^\mu(x_{n+1}, n+1) - V^\mu(x_n, n) \;\le\; -\;\lambda\, V(x_n) \;+\; O(\mu^2)\,(V(x_n)+1).
$$

Using (4.36), replacing $V(x_n)$ in the last line above by $V^\mu(x_n, n)$, taking expectation, and iterating on the resulting inequality, we arrive at

$$
\begin{aligned}
E\, V^\mu(x_{n+1},\; n+1) \le\; (1-\lambda\,\mu)^n\, V^\mu(x_0, 0) \quad &+ O(\mu^2) \sum_{k=0}^{n} (1-\lambda\,\mu)^k \\
&+\; O(\mu^2) \sum_{k=0}^{n} (1-\lambda\,\mu)^{n-k}\, V^\mu(x_k, k).
\end{aligned}
\tag{4.39}
$$

Note that there is an N_μ such that for all $n \ge N_\mu$, we can make $(1-\lambda\,\mu)^n\, EV^\mu(x_0,0) \le O(\mu)$. In addition, $\sum_{k=0}^{n} (1-\lambda\,\mu)^k\, O(\mu^2) = O(\mu)$ for all $n \le O(1/\mu)$. Using the estimates in the above paragraph, an application of the Gronwall's inequality yields that

$E\ V^{\mu}(x_{n+1}, n+1) \ \leq \ O(\mu)$. Using (4.36) again in the estimate above, we obtain $E\ V(x_n) \leq O(\mu)$. The desired result thus follows. □

Define

$$U_n = \frac{x_n - x_c}{\sqrt{\mu}} \ \text{ and } \ \widetilde{U}_n^i = \frac{\widetilde{x}_n^i - x_c}{\sqrt{\mu}}. \tag{4.40}$$

We assume that the following assumption holds.

(A3) (i) For each $\ell \ \in \ \mathcal{M}$ and each ζ, $\widehat{W}_n(\cdot, \ell, \zeta)$ has continuous partial derivatives with respect to x up to the second order and $\widehat{W}_{n,xx}(\cdot, \ell, \zeta)$ is uniformly bounded, where $\widehat{W}_{n,x}(\cdot)$ and $\widehat{W}_{n,xx}(\cdot)$ denotes the first and second partial derivatives with respect to x. The $\{\widehat{W}_n(x_c, \ell, \widetilde{\zeta}_n^i\}$ and $\{\widehat{W}_{n,x}(x_c, \ell, \widetilde{\zeta}_n^i)\}$ are bounded and uniform mixing sequences with the mixing measure satisfying $\sum_k \psi^{1/2}(k) < \infty$. (ii) For a sequence of indicator functions $\{\chi_j(A)\}$ where A is any measurable set with respect to $\{\alpha_k, Y_{k-1}^i : i \leq r, k \leq j\}$, $\sum_{j=m}^{m+n-1}\{[M_j(\ell) - \overline{M}(\ell)]x_c\}\chi_j(A)/\sqrt{n} \to 0$ in probability uniformly in m. (iii) The averaging conditions in (A2) hold with fixed m replaced by $N_i(N_\mu)$. (iv) The sets $\{\widetilde{\xi}_n^i\}$ and $\{\widetilde{\zeta}_j^i\}$ for $i = 1 \ldots, r$ are mutually independent.

Note that (A3) (i) implies that we can locally linearize $\widehat{W}_n(\cdot)$ around x_c,

$$\begin{aligned}
&\widehat{W}_n(x, \ell, \zeta) = \widehat{W}_n(x_c, \ell, \zeta) + \widehat{W}_{n,x}(x_c, \ell, \zeta)(x - x_c) + O(|x - x_c|^2), \\
&\frac{1}{n}\sum_{j=m}^{m+n-1} E_m \widehat{W}_{j,x}(x_c, \ell, \widetilde{\zeta}_j^i) \to 0 \ \text{ in probability}, \\
&\sum_{k=n}^{\infty} \ | \ E_n \ \widehat{W}_k(x_c, \ell, \widetilde{\zeta}_k^i)| \ < \infty,
\end{aligned}$$

Condition (A3) (ii) is a technical condition similar to [109, (A1.5) on p. 318]. Recall that $\varepsilon \ = \mu$. It is a requirement on the rates of local average for the sequence $\{M_j(\ell)\}$ It is readily

verified that

$$\begin{aligned} U^i_{n+1} = U^i_n \ &+\mu[M_n(\alpha_n)\widetilde{U}^i_n]^i I^i_{n+1} + \mu[\widehat{W}_{n,x}(x_c,\alpha_n,\zeta^i_n)\widetilde{U}^i_n]^i I^i_{n+1} \\ &+\sqrt{\mu}[W_n(\alpha_n)\xi^i_n]^i I^i_{n+1} + \sqrt{\mu}[\widehat{W}_n(x_c,\alpha_n,\zeta^i_n)]^i I^i_{n+1} \\ &+\sqrt{\mu}\{[M_n(\alpha_n)-\overline{M}(\alpha_n)]x_c\}^i I^i_{n+1} + O(\mu^{3/2})O(|\widetilde{U}^i_n|^2). \end{aligned} \tag{4.41}$$

To proceed, define $U^\mu(t) = U_n$ for any $t \in [\mu\,(n - N_\mu),\ \mu\,(n-N_\mu)+\ \mu)$. Under suitable conditions, we show that $\{U^\mu(\cdot)\}$ converges weakly to a switching diffusion process. First note that by (A3) (ii),

$$\begin{aligned} &\sqrt{\mu}\sum_{j=t/\mu}^{(t+s)/\mu-1}\{[M_j(\alpha_j)-\overline{M}(\alpha_j)]x_c\}^i I^i_{j+1} \\ &= \sqrt{s}\frac{1}{\sqrt{s/\mu}}\sum_{\ell\in\mathcal{M}}\sum_{j=t/\mu}^{(t+s)/\mu-1}\{[M_j(\ell)-\overline{M}(\ell)]x_c\}^i I_{\{\alpha_j=\ell\}} I^i_{j+1} \\ &\to 0 \text{ as } \ \mu\to 0 \ \text{ uniformly in } t\in[0,T]. \end{aligned}$$

Observe that by virtue of Theorem 4.10, $\{U_n:\ n\ \geq\ N_\mu\}$ is tight. It yields that

$$\begin{aligned} U^{\mu,i}(t+s)-U^{\mu,i}(t) = \mu &\sum_{j=t/\mu}^{(t+s)/\mu-1}[M_j(\alpha_j)\widetilde{U}^i_j]^i I^i_{j+1} \\ &+\mu\sum_{j=t/\mu}^{(t+s)/\mu-1}[\widehat{W}_{j,x}(x_c,\alpha_j,\zeta^i_j)\widetilde{U}^i_j]^i I^i_{j+1} \\ &+\sqrt{\mu}\sum_{j=t/\mu}^{(t+s)/\mu-1}[\widehat{W}_j(x_c,\alpha_j,\zeta^i_j)]^i I^i_{j+1} \\ &+\sqrt{\mu}\sum_{j=t/\mu}^{(t+s)/\mu-1}[W_j(\alpha_j)\xi^i_j]^i I^i_{j+1} + o(1), \end{aligned} \tag{4.42}$$

where $o(1)\to\ 0$ in probability. The $o(1)$ is obtained by use of the last line of (4.41), (A3), and Theorem 4.10. Using the methods presented for analyzing $x^\mu(\cdot)$, we obtain the following lemma, whose details are omitted.

Lemma 4.11. *$\{U^\mu(\cdot), \alpha^\mu(\cdot)\}$ is tight in $D([0,T] : \mathbb{R}^r \times \mathcal{M})$.*

Next, for $n > 0$, and each $\iota \in \mathcal{M}$, define

$$\begin{aligned}
&B^{\mu,i,\iota}_{n,1} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [\widehat{W}(x_c, \iota, \zeta^i_j)]^i) I^i_{j+1},\ B^{\mu,i,\iota}_{n,2} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [W_j(\iota)\xi^i_j]^i I^i_{j+1},\\
&B^{\mu,i,\iota}_1(t) = B^{\mu,i,\iota}_{n,1},\ B^{\mu,i,\iota}_2(t) = B^{\mu,i,\iota}_{n,2} \ \text{ for } t \in [\mu n, \mu n + \mu),\\
&\widetilde{B}^{\mu,i,\iota}_{n,1} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [\widehat{W}(x_c, \iota, \widetilde{\zeta}^i_j)]^i,\ \widetilde{B}^{\mu,i,\iota}_{n,2} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} = [W_j(\iota)\widetilde{\xi}^i_j]^i,\\
&\widetilde{B}^{\mu,i,\iota}_1(t) = \widetilde{B}^{\mu,i,\iota}_{n,1},\ \widetilde{B}^{\mu,i,\iota}_2(t) = \widetilde{B}^{\mu,i,\iota}_{n,2} \ \text{ for } t \in [\mu n, \mu n + \mu).
\end{aligned} \tag{4.43}$$

Define also

$$\begin{aligned}
&Z^{\mu,i}_n = \mu \sum_{j=N_\mu}^{N_\mu+n-1} Y^i_j,\ Z^{\mu,i}(t) = Z^{\mu,i}_n, \ \text{ for } t \in [\mu n, \mu n + \mu),\\
&b^{\mu,i}_{n,1} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [\widehat{W}(x_c, \alpha_j, \zeta^i_j)]^i) I^i_{j+1},\ b^{\mu,i}_{n,2} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [W_j(\alpha_j)\xi^i_j]^i I^i_{j+1},\\
&b^{\mu,i}_1(t) = b^{\mu,i}_{n,1},\ b^{\mu,i}_2(t) = b^{\mu,i}_{n,2} \ \text{ for } t \in [\mu n, \mu n + \mu)\\
&\widetilde{b}^{\mu,i}_{n,1} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [\widehat{W}(x_c, \widetilde{\alpha}_{\tau^i_j}, \widetilde{\zeta}^i_j)]^i,\ b^{\mu,i}_{n,2} = \sqrt{\mu} \sum_{j=N_\mu}^{N_\mu+n-1} [W_j(\widetilde{\alpha}_{\tau^i_j})\xi^i_j]^i,\\
&\widetilde{b}^{\mu,i}_1(t) = \widetilde{b}^{\mu,i}_{n,1},\ \widetilde{b}^{\mu,i}_2(t) = \widetilde{b}^{\mu,i}_{n,2} \ \text{ for } t \in [\mu n, \mu n + \mu).
\end{aligned} \tag{4.44}$$

Using similar methods of the martingale averaging as in Theorem 4.6, we can show that $B^{\mu,i,\iota}_l(\cdot)$ converges weakly to $B^{i,\iota}_l(\cdot) = \widetilde{B}^{i,\iota}_l((Z^i(\cdot))^{-1})$ for $l = 1, 2$. It is also easy to see that $B^{i,\iota}_1(\cdot)$ and $B^{i,\iota}_2(\cdot)$ are independent.

Theorem 4.12. *Under* (A1)–(A3)*, there are independent standard Brownian motions $w_{i,1}(\cdot)$ and $w_{i,2}(\cdot)$ for $i \le r$ such that the limits $U^i(\cdot)$, $i \le r$ satisfy*

$$dU^i = \frac{[\overline{M}(\alpha(t))U]^i}{\pi^i(x_c, \alpha(t))}dt + \frac{[\sigma^i_1(\alpha(t))\ dw_{i,1}(t) \ + \ \sigma^i_1(\alpha(t))dw_{i,2}(t)]}{\sqrt{\pi^i(x_c, \alpha(t))}} \quad , i \le r. \tag{4.45}$$

Proof. The proof is similar in spirit to that of Theorem 4.6. So we will only point out the distinct features. Using the well-known results for mixing processes (see [97] and [100]), it is easily seen that $\widetilde{B}_1^{\mu,i,\iota}(\cdot)$ and $\widetilde{B}_2^{\mu,i,\iota}(\cdot)$ converge weakly to Brownian motions $\widetilde{B}_1^{i,\iota}(\cdot)$ and $\widetilde{B}_2^{i,\iota}(\cdot)$, with covariance $(\sigma_1^i(\iota))^2\ t$ and $(\sigma_2^i(\iota))^2 t$, respectively. It is also easy to see that $B_1^{i,\iota}(\cdot)$ and $B_2^{i,\iota}(\cdot)$ are independent. Using the scaling argument as in the proof of Theorem 4.6, we can show that $B_l^{\mu,i,\iota}(\cdot)$ converges weakly to $B_l^{i,\iota}(\cdot) = \widetilde{B}_l^{i,\iota}((Z^i(\cdot))^{-1})$ for $l = 1, 2$. We need to prove the independence of the limit Brownian motions. Here we use an argument similar to [108, the last few lines of p. 239 and the first few lines of p. 240]. Let $K_\mu \to \infty$ such that $\sqrt{\mu}K_\mu \to 0$ and let $N_i^\mu = N_i(N_\mu)$. We work with

$$\widehat{B}_1^{\mu,i,\iota}(t) = \sqrt{\mu} \sum_{j=N_i^\mu+K_\mu}^{N_i^\mu+K_\mu-1} [\widehat{W}(x_c, \iota, \widetilde{\zeta}_j^i)]^i,\ \widehat{B}_{n,2}^{\mu,i,\iota} = \sqrt{\mu} \sum_{j=N_i^\mu+K_\mu}^{N_i^\mu+K_\mu-1} [W_j(\iota)\widetilde{\xi}_j^i]^i.$$

For simplicity of notation, we take $r = 2$. We shall show that the limits of $\widehat{B}^{\mu,i,\iota}(\cdot)$ are independent. Denote $I_{mn}^\mu = I_{\{N_1^\mu=n\}}\ I_{\{N_2^\mu=m\}}$. For any bounded and continuous function $H_1(\cdot)$ and $H_2(\cdot)$, we have

$$\begin{aligned}
&EH_1(\widehat{B}_1^{\mu,1,\iota}(t)H_2(\widehat{B}_1^{\mu,2,\iota}(t)) \\
&= \sum_{n,m} EH_1(\sqrt{\mu} \sum_{j=n+k}^{n+k+(t/\mu)} [\widehat{W}(x_c, \iota, \zeta_j^1)]^1)H_2(\sqrt{\mu} \sum_{j=m+k}^{m+k+(t/\mu)} [\widehat{W}(x_c, \iota, \zeta_j^1)]^2)I_{mn}^\mu \\
&= \sum_{n,m} EH_1(\sqrt{\mu} \sum_{j=n+k}^{n+k+(t/\mu)} [\widehat{W}(x_c, \iota, \zeta_j^1)]^1)EH_2(\sqrt{\mu} \sum_{j=m+k}^{m+k+(t/\mu)} [\widehat{W}(x_c, \iota, \zeta_j^1)]^2)EI_{mn}^\mu + o(1),
\end{aligned}$$

where $o(1) \to 0$ as $\mu \to 0$. This together with the arbitrariness of k and the weak convergence implies that the independence of the limit Brownian motions. Likewise, we can show the independence of the limit Brownian motions associated with $\widetilde{B}_2^{\mu,i,\iota}(\cdot)$.

We can then show

$$\begin{aligned}\widetilde{b}_{n,1}^{\mu,i} &= \sqrt{\mu}\sum_{j=N_\mu}^{N_\mu+n-1}[\widehat{W}_{\tau_j^i}(x_c,\widetilde{\alpha}_{\tau_j^i},\widetilde{\zeta}_j^i)]^i \\ &= \sqrt{\mu}\sum_{\iota\in\mathcal{M}}\sum_{j=N_\mu}^{N_\mu+n-1}[\widehat{W}\tau_j^i(x_c,\iota,\widetilde{\zeta}_j^i)]^i I_{\{\widetilde{\alpha}_{\tau_j^i}=\iota\}}.\end{aligned}$$

Choose m_μ, δ_μ etc. as in the convergence proof of the algorithm. Then

$$\widetilde{b}_1^{\mu,i}(t+s)-\widetilde{b}_1^{\mu,i}(t)=\sum_{\iota\in\mathcal{M}}\sqrt{\delta_\mu}\frac{1}{\sqrt{m_\mu}}\sum_{l=t/\delta_\mu}^{(t+s)/\delta_\mu}\sum_{j=lm_\mu}^{lm_\mu+m_\mu-1}[\widehat{W}_{\tau_j^i}(x_c,\iota,\widetilde{\zeta}_j^i)]^i I_{\{\widetilde{\alpha}_{\tau_j^i}=\iota\}}. \tag{4.46}$$

Let $\mu\, l\, m_\mu \to u$. Then for all $lm_\mu \le j \le lm_\mu + m_\mu - 1$, $\mu\, k \to u$. Using the weak convergence of $\widetilde{\alpha}^\mu(\cdot)$ to $\alpha(\cdot)$ and the Skorohod representation, the limit in the last line of (4.46) is the same as that of

$$\sum_{\iota\in\mathcal{M}}\sqrt{\delta_\mu}\frac{1}{\sqrt{m_\mu}}\sum_{l=t/\delta_\mu}^{(t+s)/\delta_\mu}\sum_{j=lm_\mu}^{lm_\mu+m_\mu-1}[\widehat{W}_{\tau_j^i}(x_c,\iota,\widetilde{\zeta}_j^i)]^i I_{\{\alpha(u)=\iota\}}.$$

Thus, the limit of (4.46) is given by

$$\begin{aligned}\widetilde{b}_1^i(t+s)-\widetilde{b}_1^i(t) &= \sum_{\iota\in\mathcal{M}}\int_t^{t+s}\sigma_1^i(\iota)I_{\{\alpha(u)=\iota\}}dw_{i,1}(u) \\ &= \int_t^{t+s}\sigma_1^i(\alpha(u))dw_{i,1}(u),\end{aligned} \tag{4.47}$$

where $w_{i,1}(\cdot)$ is a standard Brownian motion. Likewise, $\widetilde{b}_2^{\mu,i}(\cdot)$ converges to a switched Brownian motion in the sense that $\widetilde{b}_2^i(t) = \int_0^t \sigma^i(\alpha(u))dw_{i,2}(u)$. Thus $\widetilde{b}^{\mu,i}(\cdot) = \widetilde{b}_1^{\mu,i}(\cdot) + \widetilde{b}_2^{\mu,i}(\cdot)$

converges weakly to $\widetilde{B}^i(\cdot)$ such that

$$\widetilde{b}^i(t) = \int_0^t [\sigma_1^i(\alpha(u))dw_{i,1}(u) + \sigma_2^i(\alpha(u))dw_{i,2}(u)].$$

The last step is to combine the above estimates together with the independence of the limit Brownian motions established together with a scaling argument as in the proof of Theorem 4.6. A few details are omitted. □

Remark 4.13. In view of the independence of the Brownian motions $w_{i,1}(\cdot)$ and $w_{i,2}(\cdot)$, there is a standard Brownian motion $w_i(\cdot)$ such that the switching diffusion (4.45) may also be written in an equivalent form as

$$dU^i = \frac{[\overline{M}(\alpha(t))U]^i}{\pi^i(x_c, \alpha(t))}dt + \widehat{\sigma}^i(\alpha(t))dw_i(t),\ i \le r, \tag{4.48}$$

where

$$[\widehat{\sigma}^i(\ell)]^2 = \frac{[\sigma_1^i(\ell)]^2 + [\sigma_2^i(\ell)]^2}{\pi^i(x_c, \ell)},\quad \ell \in \mathcal{M},\ i \le r.$$

4.6 Slowly Varying ($\varepsilon \ll \mu$) and Rapidly Varying ($\mu \ll \varepsilon$) Markov Chains

This section is divided into two subsections. One of them is concerned with slowly varying Markov chains ($0 < \varepsilon \ll \mu$), whereas the other treats rapidly switching processes ($0 < \mu \ll \varepsilon$).

4.6.1 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 step size of the algorithm (4.9). 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 4.4 still holds. We next use these to analyze algorithm (4.9). As in the previous case, we can prove $\sup_{0\le n\le O(1/\mu)} E|x_n^i|^2 < \infty$. 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 $\{x^{\mu,i}(\cdot)\}$ is tight in $D([0,T],\, \mathbb{R})$. We proceed to characterize its limit. The analysis is similar to that Theorem 4.6, so we will omit most of the details.

The idea behind is that since the Markov chain is slowly varying. The the parameter is almost a constant. Since $\widetilde{\alpha}_0 = \sum_{\iota=1}^{m_0} \iota\, I_{\{\widetilde{\alpha}_0=\iota\}}$, we obtain the desired result with $[\overline{M}(\iota)x(u)]^i/\pi(x(u),\iota)$ in (4.17) replaced by $\sum_{\iota=1}^{m_0} p_\iota\, [\overline{M}(\iota)x(u)]^i/\pi^i(x(u),\iota)$. We summarize the discussions above into the following result.

Theorem 4.14. *Assume the conditions of Theorem* 4.6 *with the modification that the step-size in* (4.9) *satisfies* $\varepsilon = \mu^2$. *Then* $x^\mu(\cdot)$ *converges weakly to* $x(\cdot)$, *which is a solution of the ordinary differential equation*

$$\frac{dx^i(t)}{dt} = \sum_{\iota=1}^{m_0} p_\iota\, \frac{[\overline{M}(\iota)x(t)]^i}{\pi^i(x(t),\iota)}. \tag{4.49}$$

In addition, 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* $\delta > 0$, $\lim_{\mu\to 0} P(|x^\mu(\cdot + t_\varepsilon) - x_c| \ge \delta) = 0$.

Remark 4.15. To carry out the error analysis, furthermore, we define x_c and U_n as before and show that $\{U_n : n \ge N_\mu\}$ is tight. Letting $U^\mu(t)$ be a piecewise constant interpolation of U_n on $t \in [(n-N_\mu)\mu, (n-N_\mu)\mu+\mu)$, similar to Remark 4.13, then $U^\mu(\cdot)$ converges weakly

to $U(\cdot)$ such that $U(\cdot)$ is the solution of the stochastic differential equation

$$dU^i = \sum_{\iota=1}^{m_0} \frac{p_\iota [\overline{M}(\iota)U(t)]^i}{\pi^i(x_c, \iota)} dt + \widetilde{\sigma}_i(t) dw_i(t),$$

where $w_i(\cdot)$ is a standard Brownian motion and

$$[\widetilde{\sigma}_i(t)]^2 = \sum_{\iota=1}^{m_0} p_\iota \frac{[\sigma_1^i(\iota)]^2 + [\sigma_2^i(\iota)]^2}{\pi^i(x_c, \iota)},\ i \leq r.$$

Note that the interpolation of the centered and scaled sequence of errors has a diffusion limit in which the drift and diffusion coefficients are averaged out with respect to the initial probability distribution.

4.6.2 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 stepsize, 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.

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

$$P\{\alpha_k = j | \alpha_{lm_\mu}\} = \Xi_{ij}(\varepsilon\ lm_\mu, \varepsilon k) + O(\varepsilon + \exp(-\kappa))$$

In view of (4.16) and noting $\varepsilon = \mu^{1/2}$ and irreducibility of Q, we have $\Xi_{ij}(\varepsilon\ lm_\mu, \varepsilon k) = \nu_j + O\left(\exp\left(-\kappa_0 \frac{k - lm_m u}{\sqrt{\mu}}\right)\right)$, 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. Thus we obtain the limit ordinary differential equation.

Theorem 4.16. *Assume the conditions of Theorem* 4.6 *with the modification that the step-size in* (4.9) *satisfies* $\varepsilon = \mu^{1/2}$. *Then* $x^\mu(\cdot)$ *converges weakly to* $x(\cdot)$, *which is a solution of the ordinary differential equation*

$$\frac{dx^i(t)}{dt} = \sum_{\iota=1}^{m_0} \nu_\iota \frac{[\overline{M}(\iota)x(t)]^i}{\pi^i(x(t),\iota)}. \tag{4.50}$$

In addition, for any $t_\mu \to \infty$ *as* $\mu \to 0$ *and for any* $\delta > 0$, $\lim_{\mu\to 0} P(|x^\mu(\cdot + t_\mu) - x_c| \geq \delta) = 0$.

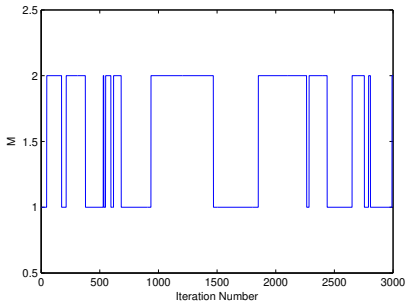
Remark 4.17. Concerning the errors, for the fast changing Markov chain case, within a very short period of time, the system is replaced by an average with respect to the stationary distribution of the Markov chain. For the error analysis, furthermore, we may define x_c U_n as in (4.40) and show that $\{U_n : n \geq N_\mu\}$ is tight. Letting $U^\mu(t)$ be the piecewise constant interpolation of U_n on $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$, similar to Remark 4.13, then $U^\mu(\cdot)$ converges weakly to $U(\cdot)$ such that $U(\cdot)$ is the solution of the stochastic differential equation

$$dU^i = \sum_{\iota=1}^{m_0} \frac{\nu_\iota[\overline{M}(\iota)U(t)]^i}{\pi^i(x_c,\iota)}dt + \overline{\sigma}_i(t)dw_i(t),$$

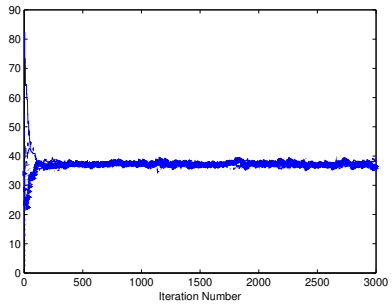
where $w_i(\cdot)$ is a standard Brownian motion and

$$[\overline{\sigma}_i(t)]^2 = \sum_{\iota=1}^{m_0} \nu_\iota \frac{[\sigma_1^i(\iota)]^2 + [\sigma_2^i(\iota)]^2}{\pi^i(x_c,\iota)},\ i \leq r.$$

Remark 4.18. As was mentioned, for convenience of presentation, we chose $\varepsilon = \mu^2$ and $\varepsilon = \sqrt{\mu}$ for the slowly varying and fast varying cases. The specific forms of μ and ε enable us to simplify the presentation. The convergence results remain essentially the same for the



(a) Markov chain sample paths



(b) System state trajectories: Values of x_n

Figure 10: Trajectories of the case $\varepsilon = \mu = 0.02$: (Horizontal axes-discrete time or iteration numbers)

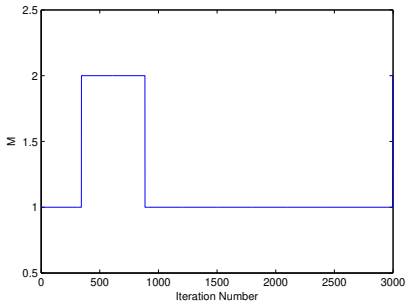
general cases $\mu/\varepsilon \to 0$ and $\mu/\varepsilon \to \infty$.

4.7 Illustrative Examples

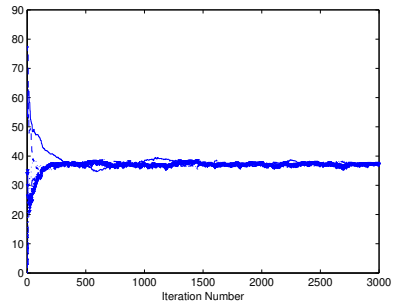
This section presents several simulation examples. We call $(x_n - x_c)'(x_n - x_c)$ the consensus error variance at time n.

Example 4.19. Suppose that the Markov chain $\widetilde{\alpha}_n$ has only 2 states, i.e., $\mathcal{M} = \{1, 2\}$. The transition probability matrix is $P^\varepsilon = I + \varepsilon\, Q$ with Q given by $\begin{pmatrix} -0.4 & 0.4 \\ 0.3 & -0.3 \end{pmatrix}$. For a given system of 5 subsystems, suppose the link gains are $G_1 = \text{diag}(1, 0.3, 1.2, 4, 7, 10)$ 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 $\eta = 37.2$ ($x_c = \eta\, \mathbb{1}$). Initial consensus error is $(x_0 - x_c)'(x_0 - x_c) = 3356.8$. Take $\varepsilon = 0.02$ and step size $\mu = \varepsilon = 0.02$. The updating algorithm runs for 3000 steps, and the stopped consensus error variance is $(x_{3000} - x_c)'(x_{3000} - x_c) = 8.0166$. In Figure 10, we plot the Markov chain state trajectories and the system state trajectories.

Example 4.20. 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

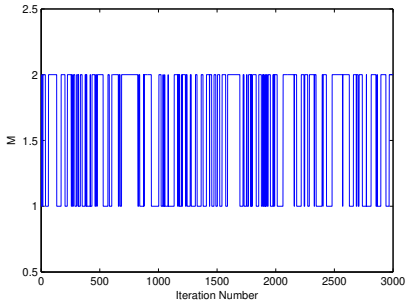


(a) Markov chain sample paths

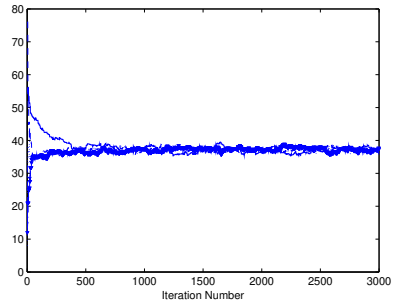


(b) System state trajectories: Values of x_n

Figure 11: Slowly varying Markov parameter $\mu = 0.02$ and $\varepsilon = \mu^2$: (Horizontal axes-discrete time or iteration numbers)



(a) Markov chain sample paths



(b) System state trajectories: Values of x_n

Figure 12: Fast varying Markov parameter $\mu = 0.02$ and $\varepsilon = \sqrt{\mu}$: (Horizontal axes-discrete time or iteration numbers)

$\mu = 0.02$. The numerical results are shown in Figure 11. From the trajectory of the Markov chain, there is only one switching taking place in the first 1000 iterations. The convergence of the consensus is also demonstrated.

Example 4.21. Here we consider the fast changing Markov $\mu \ll \varepsilon$. Specifically, we take $\mu = \varepsilon^2$ with $\mu = 0.02$. The corresponding trajectories plotted in Figure 12. The frequent Markov switching is clearly seen.

4.8 Further Remarks

For convenience and notational simplicity, we have used the current setup. Several extensions and generalizations can be carried out. 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 [109, Sections 6.6 and 8.4].

To conclude, this chapter provided a class of asynchronous stochastic approximation algorithms for consensus type of problems with randomly-switching topologies. This study extended the arenas for consensus type control problems to randomly time-varying dynamics of networked systems.

CHAPTER 5 Concluding Remarks

In this dissertation, In this dissertation, we present three applications of stochastic approximation methods. In Chapter 2, we considered a general form of PSO algorithms using a stochastic approximation scheme. Different from the existing results in the literature, we have used weaker assumptions and obtained more general results without depending on empirical work. In addition, we obtained rates of convergence for the PSO algorithms for the first time. In Chapter 3, we considered a class of stochastic approximation problems with regime switching modulated by discrete-time Markov chain. In Chapter 4, we provided a class of asynchronous stochastic approximation algorithms for consensus type of problems with randomly-switching topologies. As a rapidly expanding discipline, stochastic approximation involves a lot of techniques that go far beyond the traditional approaches. It has given impetus, not only to the applications of applied probability and stochastic processes, but also to other areas of science and engineering. Applications of stochastic methods are growing at an increasing rate. To inherit the past and to usher in the future, we perceive unprecedented challenges and opportunities for the development of stochastic approximation methods and applications in the future.

REFERENCES

[1] M. B. Nevelson and R. Z. Khasminskii, Stochastic Approximation and Recursive Estimation, Amer. Math. Soc., Providence, RI, 1976, Translation of Math. Monographs, Vol. 47.

[2] L. Ljung, Analysis of recursive stochastic algorithms, IEEE Trans. Autom. Control, AC-22:551-575, 1977.

[3] L. Ljung, On positive real transfer functions and the convergence of some recursive schemes, IEEE Trans. Autom. Control, AC-22:539-551, 1977.

[4] H. J. Kushner and D. S. Clark, Stochastic Approximation for Constrained and Unconstained Systems, Springer-Verlag, New York, 1978.

[5] M.A. Abido, "Particle swarm optimization for multimachine power system stabilizer design", in *Proc. Power Eng. Soc. Summer Meeting*, 2001, pp. 1346-1351.

[6] M. R. AlRashidi, M. E. El-Hawary, "A Survey of Particle Swarm Optimization Applications in Electric Power Systems", *IEEE Trans. Evolutionary Comp.*, vol. 13, no. 4, pp. 913-918, 2009,

[7] Quan Yuan, George Yin, Analyzing Convergence and Rates of Convergence of Particle Swarm Optimization Algorithms Using Stochastic Approximation Methods. to appear in July 2015 as a full paper in *IEEE Transactions on Automatic Control*.

[8] George Yin, Quan Yuan, Le Yi Wang, Asynchronous Stochastic Approximation Algorithms for Networked Systems: Regime-Switching Topologies and Multiscale Structure. *SIAM Multiscale Model Simulation.* 2013, 11(3): 813–839.

[9] Quan Yuan, Zhixin Yang, On the performance of a hybrid genetic algorithm in dynamic environments. *Applied Mathematics and Computation.* 2013, 219: 11408–11413.

[10] T. Beielstein, K.E. Parsopoulos, and M.N. Vrahatis, "Tuning pso parameters through sensitivity analysis", in: *Technical Report, Reihe Computational Intelligence CI 124/02*, Dept. Computer Sci., Univ. of Dortmund, 2002.

[11] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[12] P. Billingsley, *Convergence of Probability Measures*, J. Wiley, New York, NY, 1968.

[13] B. Brandstäer and U. Baumgartner, "Particle swarm optimization Mass-spring system analogon", *IEEE Trans. Magn.*, vol. 38, no. 2, pp. 997-1000, 2002.

[14] D. Bratton, and J. Kennedy, "Defining a standard for particle swarm optimization", in *Proc IEEE Swarm Intell. Symp*, 2007, pp. 120-127.

[15] X. Chen and Y. Li, "A modified PSO structure resulting in high exploration ability with convergence guaranteed". *IEEE Trans. Syst. Man Cybern. B Cybern.*, vol. 37, no. 5, pp. 1271-1289, 2007.

[16] M. Clerc and J. Kennedy, "The particle swarm: explosion, stability, and convergence in a multidimensional complex space". *IEEE Trans. Evolut. Comput.*, vol. 6, no. 1, pp. 58-73, 2002.

[17] Jr. E.F. Costa, P.L.C Lage, and Jr. E.C. Biscaia. "On the numerical solution and optimization of styrene polymerization in tubular reactors". *Comput. Chem. Eng.*, vol. 27, no. 11, pp. 1591-1604, 2003.

[18] H.M. Emara and H.A. Fattah, "Continuous swarm optimization technique with stability analysis", in *Proc. Amer. Control Conf.*, 2004, vol. 3, pp. 2811-2817.

[19] M. Jiang, Y.P. Luo, and S.Y. Yang. "Stochastic Convergence Analysis and Parameter Selection of the Standard Particle Swarm Optimization Algorithm". *Inf. Process Lett.*, vol. 102, no. 1, pp. 8-16, 2007.

[20] Enwen Zhu, Quan Yuan, P-th Moment Exponential Stability of Stochastic Recurrent Neural Networks with Markovian Switching. *Neural Processing Letters.* (2013), 1–14.

[21] Quan Yuan, Zhiqing He, A property of eigenvalue bounds for a class of symmetric tridiagonal interval matrices, *Numerical Linear Algebra with Applications.* 2010, 233: 1083–1090.

[22] Quan Yuan, Feng Qian, Wenli Du, A Hybrid Genetic Algorithm with the Baldwin Effect, Information Sciences, *Information Sciences.* 2010, 180: 640–652.

[23] C.F. Juang, "A hybrid of genetic algorithm and particle swarm optimization for recurrent network design", *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 2, pp. 997-1006, 2004.

[24] S. Kamisetty, J. Garg, J.N. Tripathi, and J. Mukherjee. "Optimization of Analog RF Circuit Parameters Using Randomness in Particle Swarm Optimization". in: *WICT*, pp. 274-278, 2011.

[25] Yu.M. Kaniovskii, "Limit distribution of processes of stochastic approximation type when the regression function has several roots", (in Russian) *Dokl. Akad. Nauk SSSR* **301** vol. 6, pp. 1308-1309, 1988.

[26] J. Kennedy, R.C. Eberhart, "Particle swarm optimization", in: *Proc. IEEE Conf. on Neural Networks, IV*, Piscataway, NJ, 1995, pp. 1942-1948.

[27] S. Kiranyaz, T. Ince, and M. Gabbouj. "Stochastic Approximation Driven Particle Swarm Optimization". in: *IIT'09*, 2009, pp. 40-44.

[28] G. Kovács, A. Groenwold, and K. Jármai, et al., "Analysis and optimum design of fibrereinforced composite structures", *Struct. Multidiscip. Opti.*, vol. 28, no. 2-3, pp. 170-179, 2004.

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

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

[31] P. L'Ecuyer and G. Yin, "Budget-dependent convergence rate of stochastic approximation", *SIAM J. Optim.* vol. 8, no. 1, pp. 217-247, 1998.

[32] Y. Li and X. Chen, "Mobile robot navigation using particle swarm optimization and adaptive NN", in: *Proc. 1st Int. Conf. Nat. Comput.*, Changsha, China, Lecture Notes in Computer Science, vol. 3612. Berlin, Germany: Springer-Verlag, 2005, pp. 554-559.

[33] Quan Yuan, Feng Qian, A Hybrid Genetic Algorithm for Twice Continuously Differential NLP Problems, *Computers & Chemical Engineering*. 2010, 34: 36–41.

[34] Quan Yuan, Zhiqing He. Bounds to eigenvalues of the Laplacian on L-shaped domain by variational methods. *Journal of Computational and Applied Mathematics.* 2009, 233: 1083–1090.

[35] Huinan Leng, Zhiqing He, Quan Yuan. Computing bounds to real eigenvalues of real-interval matrices. *International Journal for Numerical Methods in Engineering.* 2008, 74: 523–530.

[36] Quan Yuan, Zhiqing He, Huinan Leng, An Evolution Strategy Method for Computing Eigenvalue Bounds of Interval Matrices, *Applied Mathematics and Computation.* 2008, 196: 257–265.

[37] H. Liu, A. Abraham, and V. Snasel. "Convergence Analysis of Swarm Algorithm". in *NaBIC2009*, pp. 1714-1719, 2009.

[38] Y. Liu and K.M. Passino, "Stable social foraging swarms in a noisy environment", *IEEE Trans. Automat. Contr.*, vol. 49, no. 1, pp. 30-44, 2004.

[39] Y. Liu, Z. Qin, and Z. Shi, "Hybrid particle swarm optimizer with line search", in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2004, vol. 4, pp. 3751-3755.

[40] B. Liu, L. Wang, Y. Jin, "An effective PSO-based memetic algorithm for flowshop scheduling". *IEEE Trans. Syst. Man. Cy. B.-Cybernetics*, vol. 37, no. 1, pp. 18-27, 2007.

[41] J.L. Fernández-Martínez, E. García-Gonzalo, "Stochastic Stability Analysis of the Linear Continuous and Discrete PSO Models". *IEEE Trans. Evol Comput.*, vol. 15, no. 3, pp. 405-423, 2011.

[42] R. Mendes, "Population Topologies and Their Influence in Particle Swarm Performance", Ph.D. Thesis, Universidade do Minho, 2004.

[43] L. Messerschmidt and A.P. Engelbrecht, "Learning to play games using a PSO-based competitive learning approach", *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 280-288, 2004.

[44] V. Miranda, H. Keko, and A.J. Duque, "Stochastic Star Communication Topology in Evolutionary Particle Swarms (EPSO)", *Inter. J. Comput. Intelligent Research*, vol. 4, no. 2, pp. 105-116, 2008.

[45] J. Nenoraite, R. Simutis, "Stocks' trading system based on the particle swarm optimization algorithm", in: *ICCS 2004*, 2004, pp. 843-850.

[46] K.E. Parsopoulos, E.I. Papageorigiou, and P.P. Groumpos, et al. "Evolutionary computation techniques for optimizing fuzzy cognitive maps in radiation therapy systems". In: *Proc. GECCO*. 2004, pp. 402-413.

[47] N.G. Pavlidis, K.E. Parsopoulos, and M.N. Vrahatis, "Computing Nash equilibria through computational intelligence methods". *J. Comput. Appl. Math.*, vol. 175, no. 1, pp. 113-136, 2005.

[48] K.E. Parsopouls and M.N. Vrahatis, "Recent approaches to global optimization problems through particle swarm optimization", *Nat. Comput.*, vol. 1, no. 2-3, pp. 235-306, 2002.

[49] M. Pedersen, "Good parameters for particle swarm optimization". *Hvass Laboratories Technical Report no HL1001*, 2010.

[50] M.E.H. Pedersen and A.J. Chipperfield, "Simplifying Particle Swarm Optimization", *Appl. Soft Computing*, vol. 10, no. 2, pp. 618-628, 2010.

[51] R. Poli, "Dynamics and Stability of the Sampling Distribution of Particle Swarm Optimisers via Moment Analysis", *J. Artif. Evol. Appl.*, 2008. doi:10.1155/2008/761459.

[52] R. Poli, C.D. Chio, and W.B. Langdon, "Exploring extended particle swarms: A genetic programming approach", in: *Proc. Conf. Genet. & Evol. Comput.*, Washington DC, 2005, pp. 169-176.

[53] T.K. Rasmussen, T. Krink, "Improved hidden Markov model training for multiple sequence alignment by a particle swarm optimization-evolutionary algorithm hybrid". Biosystems, vol. 72, no. 1-2, pp. 5-17, 2003.

[54] A. Ratnaweera, S.K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients", *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 240-255, 2004.

[55] Quan Yuan, Zhiqing He, Huinan Leng, An improvement for Chebyshev collocation method in solving certain Sturm-Liouville problems, *Applied Mathematics and Computation.* 2008, 195: 440–447.

[56] Quan Yuan, Zhiqing He, Huinan Leng, A Hybrid Genetic Algorithm for a Class of Global Optimization Problems with Box Constraints, *Applied Mathematics and Computation.* 2008, 197: 924–929.

[57] R.G. Reynolds and C.-J. Chung, "Knowledge-based self-adaption in evolutionary programming using cultural algorithms", in: *Proc. IEEE Int. Conf. Evolutionary Computation*, Indianapolis, IN, 1997, pp. 71-76.

[58] Q. Shen, J. Jiang, and C. Jiao, et al., "Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiostensin II: antagonists". *Eur. J. Pharm. Sci.*, vol. 22, no. 2-3, pp. 145-152, 2004.

[59] Y. Shi and R. Eberhart, "A modified particle swarm optimizer", in: *Proc. IEEE World Congr. Comput. Intell.*, 1998, pp. 69-73.

[60] M.F. Tasgetiren, M. Sevkli, and Y. Liang, et al., "Partical swarm optimization algorithm for permutation flowshop sequencing problem". *Lecture Notes in Comput. Sci.*, vol. 3172, pp. 382-389, 2004.

[61] Y. Tian and Y. Takane, "The inverse of any two-by-two nonsingular partitioned matrix and three matrix inverse completion problems", *Comp. & Math. Appl.*, vol. 57, no. 8, pp. 1294-1304, 2009.

[62] I.C. Trelea, "The Particle Swarm Optimization Algorithm: convergence analysis and parameter selection". *Inf. Process Lett.* vol. 85, no. 6, pp. 317-325, 2003.

[63] H. Wu, F. Sun, and Z. Sun, et al. "Optimal trajectory planning of a flexible dual-arm space robot with vibration reduction". *J. Intell. Robot. Syst.*, vol. 40, no. 2, pp. 147-163, 2004.

[64] X. Xiao, E.R. Dow, R. Eberhart, et al., "Hybrid self-organizing maps and particle swarm optimization approach". *Concurr. Comp-Pract. E.*, vol. 16, no. 9, pp. 895-915, 2004.

[65] R. Xiao, B. Li, and X. He, "The Particle Swarm: Parameter Selection and Convergence", *CCIS*, vol. 2, pp. 396-402, 2007.

[66] K. Yasuda, A. Ide, and N. Iwasaki, "Adaptive particle swarm optimization", in: *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2003, pp. 1554-1559.

[67] G. Yin, "Rates of convergence for a class of global stochastic optimization algorithms", *SIAM J. Optim.*, vol. 10, no. 1, pp. 99-120, 1999.

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

[69] M. Zambrano-Bigiarini, M. Clerc, and R. Rojas, "Standard particle swarm optimization 2011 at CEC-2013: A baseline for future PSO improvements", *IEEE Congr. Evol. Comp. (CEC)*, Cancun, 2013, pp. 2337-2344.

[70] L. Zhang, H. Yu, and S. Hu, "Optimal choice of parameters for particle swarm optimization". *J. Zhejiang Univ. Sci.* vol. 6A, no. 6, pp. 528-534, 2005.

[71] X. Zhang, L. Yu, and Y. Zheng, et al. "Two-stage adaptive PMD compensation in a 10 Gbit/s optical communication system using particle swarm optimization". *Opt. Commun.*, vol. 231, no. 1-6, pp. 233-242, 2004.

[72] W.J. Zhang and X.F. Xie, "DEPSO: Hybrid particle swarm with differential evolution operator", in: *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2003, pp. 3816-3821.

[73] H.-F. Chen, Stochastic Approximation and Its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

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

[75] A. Benveniste, M. Metivier, and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, New York, 1990.

[76] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1968.

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

[78] H. J. Kushner and D. S. Clark, Stochastic Approximation Methods for Constrained and Unconstrained Systems, Springer-Verlag, 1978.

[79] L. Ljung, Analysis of recursive stochastic algorithms, IEEE Trans. Automat. Control, 22 (1977), pp. 551-575.

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

[81] V. Hutson and J.S. Pym, Applications of Functional Analysis and Operator Theory, Academic Press, London, 1980.

[82] J.L. Doob, Stochastic Processes, Wiley Classic Library Edition, Wiley, New York, 1990.

[83] G. Yin and Q. Zhang, Singularly perturbed discrete-time Markov chains, SIAM J. Appl. Math. 61 (2000), pp. 834-854.

[84] F.C. Hoppensteadt and W.L. Miranker, Multitime methods for systems of difference equations, Stud. Appl. Math. 56 (1977), pp. 273-289.

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

[86] T.G. Kurtz, Approximation of Population Processes, SIAM, Philadelphia, 1981.

[87] G. Yin, Q. Zhang, and G. Badowski, Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states, Ann. Appl. Probab. 10 (2000), pp. 549-572.

[88] G. Yin and H. Zhang, Discrete-time Markov chains with two-time scales and a countable stste space: limit results and queueing applications, Stochastics: An international Journal of Probability and Stochastics Processes. 4 (2008), pp. 339-369.

[89] Regime Switching Stochastic Approximation Algorithms with Application to Adaptive Discrete Stochastic Optimization, SIAM J. OPTIM. 14 (2014), pp. 1187-1215.

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

[91] A. Benveniste, M. Metivier, and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, New York, 1990.

[92] G. Yin, V. Krishnamurthy, and C. Ion, Regime Switching Stochastic Approximation Algorithms with Application to Adaptive Discrete Stochastic Optimization, SIAM J. Optim. Vol. 14, No. 4, pp. 1187-1225, 2004.

[93] G. Yin and H. Zhang, Discrete-time Markov chains with two-time scales and a countable state space: limit results and queueing applications, Stochastic An Internatioal Journal of Probability and Stochastic Processes: formly Stochastic and Stochastic Reports, 80: 4, 339-369, 2008.

[94] S. Andradóttir, A global search method for discrete stochastic optimization, SIAM J. Optim., 6 (1996), pp. 513-530.

[95] S. Andradóttir, Accelerating the convergence of random search methods for discrete stochastic optimization, ACM Trans. Model. Comput. Simul., 9 (1999), pp. 349-380.

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

[97] P. Billingsley, *Convergence of Probability Measures*, J. Wiley, New York, NY, 1968.

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

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

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

[101] Simon Haykin, Digital Communications, 4th ed., John Wiley & Sons, 2001

[102] 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.

[103] 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.

[104] 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.

[105] 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.

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

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

[108] H.J. Kushner and G. Yin, Stochastic approximation algorithms for parallel and distributed processing, *Stochastics*, **22** (1987), 219–250.

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

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

[111] 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.

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

[113] Todd K. Moon, Error Correction Coding, Mathematical Methods and Algorithms, John Wiley & Sons, 2005.

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

[115] 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.

[116] 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.

[117] 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.

[118] 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.

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

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

[121] 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.

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

[123] 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.

[124] 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.

[125] 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.

[126] 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.

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

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

[129] G. Yin, L. Y. Wang, and Y. Sun, Stochastic recursive algorithms for networked systems with delay and random switching: multiscale formulations and asymptotic properities, *SIAM Multiscale Model. Simul.*, **9** (2011), 1087–1112.

[130] 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.

[131] G. Yin and Q. Zhang. Singularly perturbed discrete-time Markov chains. *SIAM J. Applied Math*, 61:833–854, 2000.

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

[133] 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.

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

[135] G. Yin and Y.M. Zhu, On w.p.1 convergence of a parallel stochastic approximation algorithm, *Probab. Eng. Inform. Sci.*, **3** (1989), 55–75.

ABSTRACT

STOCHASTIC APPROXIMATION ALGORITHMS WITH APPLICATIONS TO PARTICLE SWARM OPTIMIZATION, ADAPTIVE OPTIMIZATION, AND CONSENSUS

by

QUAN YUAN

August 2015

Advisor: Dr. George Yin

Major: Mathematics

Degree: Doctor of Philosophy

In this dissertation, we present three three problems arising in recent applications of stochastic approximation methods. In Chapter 2, we use stochastic approximation to analyze Particle Swarm Optimization (PSO) algorithm. We introduce four coefficients and rewrite the PSO procedure as a stochastic approximation type iterative algorithm. Then we analyze its convergence using weak convergence method. It is proved that a suitably scaled sequence of swarms converge to the solution of an ordinary differential equation. We also establish certain stability results. Moreover, convergence rates are ascertained by using weak convergence method. A centered and scaled sequence of the estimation errors is shown to have a diffusion limit. In Chapter 3, we study a class of stochastic approximation algorithms with regime switching that is modulated by a discrete Markov chain having countable state spaces and two-time-scale structures. In the algorithm, the increments of a sequence of occupation measures are updated using constant step size. It is demonstrated that least squares estimations from the tracking errors can be developed. Under the assumption that the adaptation rates are of the same order of magnitude as that of times-different parameter, it is

proven that the continuous-time interpolation from the iterates converges weakly to some system of ordinary differential equations (ODEs) with regime switching, and that a suitably scaled sequence of the tracking errors converges to a system of switching diffusion. This work is an extension of the work in [92]. In Chapter 4, we developed asynchronous stochastic approximation (SA) algorithms for networked systems with multi-agents and regime-switching topologies to achieve consensus control. There are several distinct features of the algorithms. (1) In contrast to the most existing consensus algorithms, the participating agents compute and communicate in an asynchronous fashion without using a global clock. (2) The agents compute and communicate at random times. (3) The regime-switching process is modeled as a discrete-time Markov chain with a finite state space. (4) The functions involved are allowed to vary with respect to time hence nonstationarity can be handled. (5) Multi-scale formulation enriches the applicability of the algorithms. In the setup, the switching process contains a rate parameter $\varepsilon > 0$ in the transition probability matrix that characterizes how frequently the topology switches. The algorithm uses a step-size μ that defines how fast the network states are updated. Depending on their relative values, three distinct scenarios emerge. Under suitable conditions, it is shown that 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, or to a system of differential equations (an average with respect to certain measure). In addition, a scaled sequence of tracking errors converges to a switching diffusion or a diffusion. Simulation results are presented to demonstrate these findings.

AUTOBIOGRAPHICAL STATEMENT

Quan Yuan

Education

Ph.D. in Applied Mathematics, Aug, 2015 (expected)
Wayne State University, Detroit, Michigan

M.A. in Mathematical Statistics, May, 2014
Wayne State University, Detroit, Michigan

B.S. in Mathematics, June 2001
East China University of Science and Technology, Shanghai, China

Awards (Selected)

2013 Outstanding Reviewer, *Applied Soft Computing*, Mar 2014

Thomas C. Rumble University Fellowship recipient, Wayne State University, August 2014 – 2015

The Paul A. Catlin Award in recognition of outstanding achievement in the Master's Program, Department of Mathematics, Wayne State University, 2013–2014

List of Publications (Selected)

1. Quan Yuan, George Yin, Analyzing Convergence and Rates of Convergence of Particle Swarm Optimization Algorithms Using Stochastic Approximation Methods. to appear in July 2015 as a full paper in *IEEE Transactions on Automatic Control*.

2. George Yin, Quan Yuan, Le Yi Wang, Asynchronous Stochastic Approximation Algorithms for Networked Systems: Regime-Switching Topologies and Multiscale Structure. *SIAM Multiscale Model Simulation.* 2013, 11(3): 813–839.

3. Quan Yuan, Zhixin Yang, On the performance of a hybrid genetic algorithm in dynamic environments. *Applied Mathematics and Computation.* 2013, 219: 11408–11413.

4. Enwen Zhu, Quan Yuan, P-th Moment Exponential Stability of Stochastic Recurrent Neural Networks with Markovian Switching. *Neural Processing Letters.* (2013), 1–14.

5. Quan Yuan, Zhiqing He, A property of eigenvalue bounds for a class of symmetric tridiagonal interval matrices, *Numerical Linear Algebra with Applications.* 2010, 233: 1083–1090.

6. Quan Yuan, Feng Qian, Wenli Du, A Hybrid Genetic Algorithm with the Baldwin Effect, Information Sciences, *Information Sciences.* 2010, 180: 640–652.

More information, please see my google scholar page
`http://scholar.google.com/citations?user=e1mMI_wAAAAJ&hl=en`