Sign-Regressor Adaptive Filtering Algorithms for Markovian Parameters

G. George Yin, Araz Hashemi, and Le Yi Wang

ABSTRACT

This work is devoted to analyzing adaptive filtering algorithms with the use of sign-regressor for randomly time-varying parameters (a discrete-time Markov chain). In accordance with different adaption and transition rates, we analyze the corresponding asymptotic properties of the algorithms. When the adaptation rate is in line with the transition rate, we obtain a limit of a Markov switched differential equation. When the Markov chain is slowly changing the parameter process is almost a constant, and we derive a limit differential equation. When the Markov chain is fast varying, the limit system is again a differential equation that is an average with respect to the stationary distribution of the Markov chain. In addition to the limit dynamic systems, we obtain asymptotic properties of centered and scaled tracking errors. We obtain mean square errors to illustrate the dependence on the stepsize as well as on the transition rate. The limit distributions in terms of scaled errors are studied by examining certain centered and scaled error sequences.

Key Words: Sign-regressor algorithm, regime switching, stochastic approximation, mean square error, convergence, tracking property.

I. Introduction

This paper studies asymptotic properties of a class of adaptive filtering algorithms with sign regressors and time-varying system parameters. The traditional adaptive filtering problems can be described as follows. Let $\varphi_n \in \mathbb{R}^r$ and $y_n \in \mathbb{R}$ be the measured output and reference signals, respectively. Assuming that the sequence $\{(y_n, \varphi_n)\}$ is stationary, we adjust the system parameter $\varsigma$ adaptively so that the weighted output $\varsigma' \varphi_n$ best matches the reference signal $y_n$ in the sense that a cost function $L(\varsigma)$ is minimized. To solve the problem, we construct a recursive algorithm $s_{n+1} = s_n - \mu \varphi_n (y_n - \varphi_n' s_n)$.

Results on such algorithms can be found in [16, 17, 24] among others; see also [10, 18] for related problems involving time-varying parameters and stability of product of random matrices. While most of the existing references deal with decreasing stepsize algorithms with $\mu$ replaced by $\{\mu_n\}$ where $\{\mu_n\}$ is a sequence of positive scalars satisfying $\sum_n \mu_n = \infty$, $\mu_n \to 0$ as $n \to \infty$, it is known that constant stepsize algorithms have the advantage of having tracking ability for treating time-varying parameters. Algorithm (1) is of stochastic approximation type; see [2, 4, 5, 14] among others for reference on stochastic approximation.

Adaptive filtering algorithms have played an important role in the emerging technology. In addition to many applications in adaptive control, estimation, and signal processing, [2, 4, 5, 6, 9, 13, 14, 16, 17, 18, 21, 24], our motivation stems from blind multiuser detection problems arising from CDMA/DS (Code-division multiple-access implemented with direct-sequence) wireless communications. The problem can
be recast into the form of adaptive filtering; see [11, 19, 26]. It is well known that for a fixed parameter case, the mean squares algorithms have better performance since they use much more information. Owing to the use of the sign operator, the algorithms can be easily implemented and multiplications can be replaced by simple bit shifts. As a result, it becomes appealing in various applications. In addition, in many applications such as automotive systems, communications, and system identification (see [23]), some signals are physically constrained to be binary valued, naturally introducing the sign operators. Thus understanding of key properties of such algorithms are of practical importance.

In comparison to the existing literature on signal processing, parameter estimation, and system identification, this paper focuses on asymptotic properties of a class of sign-regessor adaptive filtering algorithms. Since the sign algorithms are easy to implement and accommodate application constraints, there has been much work devoted to their features [1, 7, 8, 9, 15, 21]. However, departing from the usual route, we consider systems whose parameters are no longer constant but randomly time-varying processes modeled by discrete-time Markov chains. The randomly time-varying nature makes the problem much more difficult to treat than that of constant parameters.

Instead of (1), by taking sign componentwise, we obtain the following sign-regressor algorithm:

\[
\alpha_{n+1} = \alpha_n + \mu \text{Sgn}(\varphi_n)(y_n - \varphi_n' \alpha_n),
\]

where \( \text{Sgn}(\varphi) = (\text{sgn}(\varphi^1), \ldots, \text{sgn}(\varphi^r))' \), \( \varphi = (\varphi^1, \ldots, \varphi^r) \in \mathbb{R}^r \), and \( \text{sgn}(\varphi_i) = 1_{\{\varphi_i > 0\}} - 1_{\{\varphi_i < 0\}} \), \( i = 1, \ldots, r \), and \( \mathbb{1}_A \) is the indicator of \( A \).

Our new contributions in this paper are featured by the following distinct characteristics. First we develop sign-regressor algorithms for time-varying parameters modeled by a stochastic process. Moreover, we concentrate on multi-scale structures. Note that the sign-regressor algorithms naturally come into play due to the application requirement. The salient feature of the time-varying parameter is highlighted in the Markov structure. In our setup, there are two inherent time scales. The Markov chain’s state changes with its transition frequency described by a small parameter \( \varepsilon \). On the other hand, the stepsize of the approximation sequence is \( \mu \). Interaction between \( \varepsilon \) and \( \mu \) results multi-scale structures. We use “\( \ll \)” or “\( \gg \)”, the standard mathematical (“essentially less than or essentially greater than”) notation. That is, both \( \mu \to 0 \) and \( \varepsilon \to 0 \), but \( \mu \to 0 \) much faster than \( \varepsilon \) or \( \mu / \varepsilon \to 0 \). Likewise, the use of \( O(\cdot) \) and \( o(\cdot) \) follows standard notation.

For example, \( \varepsilon = O(\mu^{1+\delta}) \) for some \( \delta > 0 \) means that \( \varepsilon \) is the same order as that of \( \mu^{1+\delta} \), which can also be written as \( \varepsilon \ll \mu \). In accordance with the relative size, we have three cases to consider: (a) \( \mu = O(\varepsilon) \): the adaptation rate is in line with that of the transition rate of the parameter; (b) \( \mu \gg \varepsilon \): the time-varying parameter changes much slower than that of the stepsize; (c) \( \mu \ll \varepsilon \): the parameter process is fast changing. Corresponding to each of the cases, the asymptotic behavior is fundamentally different. We analyze the behavior of the algorithms and provide insight on their asymptotic properties.

The rest of the paper is arranged as follows. Section II presents the precise formulation of the problem. Section III analyzes convergence properties of the algorithms. Section IV proceeds with mean squares type estimates. Section V obtains asymptotic distributions of the algorithms and reveals the tracking ability by means of examining scaled tracking error sequences. Section VI gives numerical results, and concludes the paper with further remarks.

### II. Formulation

Let \( \{y_n\} \) be a sequence of real-valued signals representing the observations obtained at time \( n \), and \( \{\alpha_n\} \) be the time-varying true parameter process, an \( \mathbb{R}^r \)-valued random process. Suppose that

\[
y_n = \varphi_n' \alpha_n + e_n,
\]

where \( \varphi_n \in \mathbb{R}^r \) is the regression vector and \( e_n \in \mathbb{R} \) is a sequence of zero mean random variables representing the error or noise. Note that (3) is a variant of the usual linear regression model, in which, a time-varying parameter changes much slower than that of constant parameters.

(A1) There is a small parameter \( \varepsilon > 0 \) such that \( \alpha_n \) is a discrete-time homogeneous Markov chain with state space \( M = \{a_1, \ldots, a_{m_0}\} \), \( a_i \in \mathbb{R}^r \), \( i = 1, \ldots, m_0 \), and transition probability matrix given by

\[
P^\varepsilon = I + \varepsilon Q,
\]

where \( I \) is an \( \mathbb{R}^{m \times m} \) identity matrix and \( Q = (q_{ij}) \in \mathbb{R}^{m_0 \times m_0} \) is an irreducible generator (i.e., \( Q \) satisfies \( q_{ij} \geq 0 \) for \( i \neq j \) and \( \sum_{j=1}^{m_0} q_{ij} = 0 \) for each \( i = 1, \ldots, m_0 \)) of a continuous-time Markov chain.

(A2) Let \( \mathcal{F}_n \) be the \( \sigma \)-algebra generated by \( \{(\varphi_j, e_j), \alpha_j : j < n; \alpha_n\} \), and denote the conditional expectation with respect to \( \mathcal{F}_n \) by \( E_n \). The sequences...
{φn} and {cn} are independent of αn, the parameter process. In addition, \( \{ (φn, cn) \} \) is a sequence of bounded signals such that there is a stable matrix \( G \in \mathbb{R}^{r \times r} \) and \( K > 0 \) satisfying that for each \( n \), \( | \sum_{j=n}^{∞} E_n | \text{Sgn}(φ_j) φ_j - G | | ≤ K \) and \( | \sum_{j=n}^{∞} E_n | \text{Sgn}(φ_j) e_j | | ≤ K \).

**Remark II.1** Before proceeding further, let us comment on the conditions briefly. In (A1), for simplicity, we take \( Q \) to be a constant matrix. Time-varying \( Q(t) \) can be considered in [27], but the notation is more complex. For convenience, we assume the initial distribution of the Markov chain \( \alpha_n \) is given by \( P(\alpha_0 = α_i) = p_{0,i} \), independent of \( ε \) for each \( i = 1, . . . , m_0 \), where \( p_{0,i} ≥ 0 \) and \( \sum_{i=0}^{m_0} p_{0,i} = 1 \). In (A2), by a stable matrix \( G \), we mean that \( G \) is Hurwitz or all of its eigenvalues have negative real parts.

To track the parameter \( \{ α_n \} \), we construct a sequence of estimates given by (2), where \( μ > 0 \) is a small constant stepsize. This is a recursive algorithm of stochastic approximation type. Different from the traditional adaptive filtering setup, (i) sign-regressor is used and (ii) the parameter is a Markov chain. As a result, the standard results on adaptive filtering do not carry over. Special attention has to be paid to the sign operator and the randomly time-varying parameter.

The model of the signal includes a large class of practical cases. The inequalities in (A2) are essentially a mixing-type condition, which indicates that remote past and distant future are asymptotically independent. This condition enables us to work with correlated signals as long as the correlation decays sufficiently fast. Note that the distributions of the signals need not be known. The boundedness is a mild restriction; for instance, we may use a truncated Gaussian process, etc. In addition, when dealing with recursive procedures in practice, in lieu of (2) we often use a projection or truncation algorithm of the form \( s_{n+1} = \pi_H [s_n + μSgn(φ_n)(y_n - φ_n^T s_n)] \), where \( π_H \) is a projection operator and \( H \) is a bounded set. When the iterates are outside \( H \), they will be projected back to the constrained set \( H \). More discussions for projection algorithms are in [14]. On the other hand, it is possible to work with unbounded signals. That is, in lieu of the conditions in (A2), we may assume \( \{ (φ_n, cn) \} \) is a martingale difference sequence satisfying \( E|φ_n|^4 + Δ < ∞ \) and \( E|φ_n e_n|^2 + Δ < ∞ \) for some \( Δ > 0 \). Furthermore, we can also treat moving average type signals driven by martingale difference sequences; see [25] for more discussions. In what follows, for definiteness, we will deal mainly with processes satisfying (A2). Dealing with uncorrelated random sequences and the proof for the unbounded martingale difference sequence is slightly simpler. The problem that we encounter is a multi-scale one. Concerning the relative sizes of \( μ \) and \( ε \), there are several possibilities. We could have (i) \( μ = O(ε) \), (ii) \( μ \geq ε \), or (iii) \( μ \ll ε \).

### III. Asymptotic Properties: Convergence

This section is divided into three subsections in accordance with the three cases mentioned above. We use weak convergence methods to carry out the analysis. Before proceeding further, let us recall some definitions and notation for weak convergence. Let \( X_n \) and \( X \) be \( \mathbb{R}^r \)-valued random vectors. We say \( X_n \) converges weakly to \( X \) if for any bounded and continuous function \( f(·) \), \( Ef(X_n) → Ef(X) \) as \( n → ∞ \). The sequence \( \{ X_n \} \) is tight if for each \( η > 0 \), there is a compact set \( K_η \) such that \( P(X_n \notin K_η) ≥ 1 - η \) for all \( n \). The definitions of weak convergence and tightness extend to random elements in metric spaces. On a complete separable metric space, tightness is equivalent to relative compactness, which is known as the Prohorov’s Theorem. By virtue of this theorem, we are able to extract convergent subsequences once tightness is verified. In what follows, we use a martingale problem formulation to establish the desired weak convergence. This usually requires first tightness be proved and then the limit process be characterized. We refer the interested reader to [14, Chapter 7] for further details on weak convergence and related matters.

#### 3.1. Switching ODE Limit: \( μ = O(ε) \)

For simplicity, we take \( μ = ε \). To study the asymptotic properties of the sequence \( \{ s_n \} \), we take a continuous-time interpolation of the process. Define \( ζ^μ(t) = s_n, \alpha^μ(t) = α_n \) for \( t \in [nμ, nμ + μ] \). Note that this interpolated sequence has the property that its sample paths are in \( D([0, ∞) : \mathbb{R}^r) \), the space of \( \mathbb{R}^r \)-valued functions that are right continuous, have left limits, and are endowed with the Skorohod topology [14, Chapter 7]. We proceed to prove that \( ζ^μ(·) \) converges weakly to a system of randomly switching ordinary differential equations. The result is stated below. Its proof is obtained by means of establishing a number of lemmas.

**Theorem III.1** Assume that (A1) and (A2) hold. In addition, assume that there exists a constant matrix \( G \)
such that for each positive integer \(m\), as \(n \to \infty\),
\[
\frac{1}{n} \sum_{k=m}^{n+m-1} E_m \text{Sgn}(\varphi_k) \varphi'_k \to G \text{ in probability},
\]
\[
\frac{1}{n} \sum_{k=m}^{n+m-1} E_m \text{Sgn}(\varphi_k) e_k \to 0 \text{ in probability}.
\]

Then the process \((\varsigma^\mu(\cdot), \alpha^\mu(\cdot))\) converges weakly to \((\varsigma(\cdot), \alpha(\cdot))\) such that \(\alpha(\cdot)\) is a continuous-time Markov chain generated by \(Q\) and the limit process \(\varsigma(\cdot)\) satisfies the Markov switched ordinary differential equation
\[
\dot{\varsigma}(t) = G(\alpha(t) - \varsigma(t)), \quad \varsigma(0) = \varsigma_0.
\]

Note that in the classical results on adaptive filtering for uncorrelated stationary signals, one normally assumes that \(E \varphi_k \varphi'_k = \bar{G}\) a positive definite matrix. Here we are using \(\text{Sgn}\) operator and \(\text{Sgn}(\varphi)\varphi'\) is still a matrix. To prove the theorem, we first establish a number of lemmas. To begin with, we use a truncation device [14, p. 284]. Recall that \(\varsigma^N(\cdot)\) is said to be an \(N\)-truncation of \(\varsigma(\cdot)\) if \(\varsigma^N(\cdot)\) is equal to \(\varsigma(\cdot)\) upon the first exit from \(S_N = \{\varsigma \in \mathbb{R}^r : |\varsigma| \leq N\}\), the ball with radius \(N\). Define \(q^N(\cdot)\) as a truncation function that is a sufficiently smooth function and that is equal to 1 for \(\varsigma \in S_N\), and 0 for \(\varsigma \in S_{N+1}\). Then we modify algorithm (2) so that
\[
\varsigma^N_{n+1} = \varsigma^N_n + \mu \text{Sgn}(\varphi_n)(y_n - \varphi'_n s^N_n)q^N(\varsigma^N_n).
\]

Define \(\varsigma^{N,\mu}(t) = \varsigma^N_n\) for \(t \in [\mu n, \mu n + \mu]\). Our plan is that we first show the sequence \((\varsigma^{N,\mu}(\cdot), \alpha^\mu(\cdot))\) is tight. Then by Prohorov’s theorem we can extract a convergent subsequence. We show that the weak limit satisfies a switched differential equation. Finally, we let the truncation level \(N\) grow and show the untruncated sequence is also convergent in the sense of weak convergence.

**Lemma III.2** The sequence \((\varsigma^{N,\mu}(\cdot), \alpha^\mu(\cdot))\) is tight in \(D([0, \infty), \mathbb{R}^r \times \mathcal{M})\).

**Proof of Lemma** III.2. Note that the sequence \(\{\alpha^\mu(\cdot)\}\) is tight by virtue of the result of [27, Theorem 4.3]. In addition, \(\alpha^\mu(\cdot)\) converges weakly to a Markov chain generated by \(Q\). To proceed, we examine the asymptotics of the sequence \(\varsigma^\mu(\cdot)\). We have that for any \(\delta > 0\), and \(t, s > 0\) satisfying \(s \leq \delta\),
\[
E \left| \varsigma^{N,\mu}(t+s) - \varsigma^{N,\mu}(t) \right|^2 \\
\leq KE_\mu \left( \sum_{k=\mu n}^{(t+s)/\mu - 1} \text{Sgn}(\varphi_k) (\alpha_k - \varsigma^N_k) q^N(\varsigma^N_k) \right)^2 \\
+ KE_\mu \left( \sum_{k=\mu n + 1}^{(t+s)/\mu - 1} \text{Sgn}(\varphi_k) e_k \right)^2.
\]

By the boundedness of \(\alpha^\mu_k\) and \(\varsigma^{N,\mu}_k\) and the moment conditions for \(\varphi_k\) given in (A2) together with the Cauchy-Schwarz inequality, we have
\[
E_\mu \left( \sum_{k=\mu n}^{(t+s)/\mu - 1} \text{Sgn}(\varphi_k) (\alpha_k - \varsigma^N_k) q^N(\varsigma^N_k) \right)^2 \\
\leq K\mu s \sum_{k=\mu n}^{(t+s)/\mu - 1} E|\text{Sgn}(\varphi_k) e_k|^2 \leq K s^2 \leq K\delta^2.
\]

In the above and throughout the rest of the paper, \(K\) is used as a generic positive constant, whose values may be different for different appearances. Likewise, we obtain that
\[
E_\mu \left( \sum_{k=\mu n + 1}^{(t+s)/\mu - 1} \text{Sgn}(\varphi_k) e_k \right)^2 \leq K s \leq K\delta.
\]

Using (9) and (10), we arrive at
\[
\lim_{\delta \to 0} \lim_{\mu \to \infty} \sup_{0 \leq s \leq \delta} E_{t} \left[ \varsigma^{N,\mu}(t+s) - \varsigma^{N,\mu}(t) \right]^2 = 0,
\]
where \(E_{t}^{\mu}\) denotes the conditioning on \(\mathcal{F}_{t}^{\mu}\). Thus, by the criterion [12, p.47], the tightness is proved. \(\square\)

Since \((\varsigma^{N,\mu}(\cdot), \alpha^\mu(\cdot))\) is tight, it is sequentially compact. By virtue of Prohorov’s theorem, we can extract a weakly convergence subsequence. Select such a subsequence and still denote it by \((\varsigma^{N,\mu}(\cdot), \alpha^\mu(\cdot))\) for notational simplicity. Denote the limit by \((\varsigma^N(\cdot), \alpha(\cdot))\). We proceed to characterize the limit process. The result is stated in the next lemma; see [28] for the definition of operators involving a switching process.

**Lemma III.3** Assume the conditions of Lemma III.2 are satisfied and that (5) holds. Then \((\varsigma^{N,\mu}(\cdot), \alpha^\mu(\cdot))\) converges weakly to \((\varsigma^N(\cdot), \alpha(\cdot))\) that is a solution of the martingale problem with operator
\[
L_{\mu} f(\varsigma^N, a_i) = \nabla f'(\varsigma^N, a_i) G(\varphi_{i-1} - \varsigma^N) q^N(\varsigma^N) \\
+ \sum_{j=1} q_{ij} f(\varsigma^N, a_j),
\]

(11)
where \( f(\cdot, a_i) \in C_0^1 \) (\( C^1 \) function with compact support) for each \( a_i \in \mathcal{M} \).

**Proof.** For each \( t, s > 0 \),

\[
\zeta^{N,\mu}(t+s) - \zeta^{N,\mu}(t) \leq \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu \text{Sgn}(\varphi_k)\varphi_k^*(\alpha_k - \zeta_k^N)q^N(\zeta_k) + \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu \text{Sgn}(\varphi_k)e_k. \tag{12}
\]

For each bounded and continuous function \( h(\cdot) \), each \( t, s > 0 \), each positive integer \( \kappa \), and each \( i \leq t \) for \( i \leq \kappa \), denote \( h_N = h(\zeta^N(t_i), \alpha(t_i) : i \leq \kappa) \) and \( h_\mu^N = h(\zeta^{N,\mu}(t_i), \alpha^\mu(t_i) : i \leq \kappa) \). To derive the martingale limit, we need only show that for any \( f(\cdot, a_i) \in C_0^1 \),

\[
EH_N[f(\zeta^N(t+s), \alpha(t+s)) - f(\zeta^N(t), \alpha(t))] - \int_t^{t+s} L^N_1 f(\zeta^N(\tau), \alpha(\tau))d\tau = 0. \tag{13}
\]

To verify (13), we use the process indexed by \( \mu \). We treat each of the last two terms in the summand of (12) separately. Subdivide the interval with the end points \( t/\mu \) and \( (t+s)/\mu -1 \) by choosing \( m_\mu \) such that \( m_\mu \to \infty \) as \( \mu \to 0 \) but \( \delta_\mu = \mu m_\mu \to 0 \). By the smoothness of \( f(\cdot, a_i) \), it is readily seen that as \( \mu \to 0 \),

\[
EH_N^\mu[f(\zeta^{N,\mu}(t+s), \alpha^\mu(t+s)) - f(\zeta^{N,\mu}(t), \alpha^\mu(t))] \to EH_N^\mu[f(\zeta^N(t+s), \alpha(t+s)) - f(\zeta^N(t), \alpha(t))]. \tag{14}
\]

Next, using (7), we have

\[
EH_N^\mu[f(\zeta^{N,\mu}(t+s), \alpha^\mu(t+s)) - f(\zeta^{N,\mu}(t), \alpha^\mu(t))] = EH_N^\mu\left[\sum_{l_\mu=t}^{t+s} f(\zeta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu}) - f(\zeta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu})\right] + \int_t^{t+s} L^N_1 f(\zeta^N(\tau), \alpha^\mu(\tau))d\tau. \tag{15}
\]

Note that

\[
\lim_{\mu \to 0} EH_N^\mu\left[\sum_{l_\mu=t}^{t+s} f(\zeta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu}) - f(\zeta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu})\right] = \lim_{\mu \to 0} EH_N^\mu\sum_{l_\mu=t}^{t+s} \frac{\delta_\mu}{m_\mu} \sum_{k=lm_\mu+1}^{lm_\mu+m_\mu-1} \nabla f'(\zeta^N_{lm_\mu}, \alpha_{lm_\mu+m_\mu}) \times \text{Sgn}(\varphi_k)\varphi_k^*(\alpha_k - s_k^N)q^N(\zeta_k)
\]

\[
+ \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla f'(\zeta^N_{lm_\mu}, \alpha_{lm_\mu+m_\mu}) \times (s_k^N - s_k)q^N(\zeta_k)
\]

\[
+ \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla f'(\zeta^N_{lm_\mu}, \alpha_{lm_\mu+m_\mu}) \times \text{Sgn}(\varphi_k)\varphi_k^* e_k q^N(\zeta_k), \tag{16}
\]

where \( \zeta^{N,\mu} \) is a point on the line segment joining \( \zeta_{lm_\mu}^N \) and \( \zeta_{lm_\mu+m_\mu}^N \). By the continuity of \( \nabla f(\cdot, a_i) \),

\[
\lim_{\mu \to 0} EH_N^\mu\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla f'(\zeta_{lm_\mu}^N, \alpha_{lm_\mu+m_\mu}) - \nabla f'(\zeta_{lm_\mu}^N, \alpha_{lm_\mu+m_\mu}) \times (s_k^N - s_k^N)q^N(\zeta_k) = 0. \tag{17}
\]

Let \( \mu m_\mu \to \tau \). Then for all \( k \) satisfying \( lm_\mu \leq k \leq lm_\mu + m_\mu - 1 \), we have \( \mu k \to \tau \) as well. Thus, using (5),

\[
\lim_{\mu \to 0} EH_N^\mu\int_t^{t+s} \nabla f'(\zeta^N(\tau), \alpha(\tau))d\tau = \lim_{\mu \to 0} EH_N^\mu\sum_{l_\mu=t}^{t+s} \frac{\delta_\mu}{m_\mu} \sum_{k=lm_\mu+1}^{lm_\mu+m_\mu-1} \nabla f'(\zeta^N_{lm_\mu}, \alpha_{lm_\mu+m_\mu}) \times \text{Sgn}(\varphi_k)\varphi_k^* e_k q^N(\zeta_k)
\]

\[
= EH_N^\mu\int_t^{t+s} \nabla f'(\zeta^N(\tau), \alpha(\tau))G(\tau)q^N(\zeta^N(\tau))d\tau. \tag{18}
\]

By the smoothness of \( f(\cdot, a_i) \) and the continuity of a linear function \( \zeta^N(\cdot) \), we obtain

\[
\lim_{\mu \to 0} EH_N^\mu\sum_{l_\mu=t}^{t+s} \frac{\delta_\mu}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla f'(\zeta^N_{lm_\mu}, \alpha_{lm_\mu+m_\mu}) \times \text{Sgn}(\varphi_k)\varphi_k^* e_k q^N(\zeta_k)
\]

\[
= EH_N^\mu\int_t^{t+s} \nabla f'(\zeta^N(\tau), \alpha(\tau))G(\tau)q^N(\zeta^N(\tau))d\tau. \tag{19}
\]
Also by (5), we have
\[
\lim_{\mu \to 0} E h_N^s \sum_{l_{\mu}=t}^{t+s} \frac{1}{m_{\mu}} \sum_{k=l_{\mu}}^{l_{\mu}+m_{\mu}-1} \nabla f^s_N(\hat{s}_{l_{\mu},k}, \alpha_{l_{\mu}}) 
\times \text{Sgn}(\varphi_k) e_k q^N(\hat{s}^N) = 0. \tag{20}
\]
Similarly, we can show that
\[
\lim_{\mu \to 0} E h_N^N \sum_{l_{\mu}=t}^{t+s} \frac{1}{m_{\mu}} \sum_{k=l_{\mu}}^{l_{\mu}+m_{\mu}-1} [-f^N_N(\hat{s}_{l_{\mu},k}, \alpha_{l_{\mu}})] 
= E h_N \left[ \int_t^{t+s} \int \sum_{l_{\mu}=t}^{l_{\mu}+m_{\mu}-1} Q f(\hat{s}^N(r), \alpha(r)) dr \right]. \tag{21}
\]
Combining the estimates (14)–(21), the desired property (13) is established. The lemma is thus proved.

\[\square\]

**Completion of the Proof of Theorem III.1.** Next, letting \( N \to \infty \), we show that the limit of the untruncated sequence and the limit of the truncated sequence as \( N \to \infty \) are the same. The argument is similar to that of [14, pp. 249-250]; we explain the main steps below. Let \( P^0(\cdot) \) and \( P^N(\cdot) \) be the measures induced by \( \psi(\cdot) \) and \( \psi^N(\cdot) \), respectively. Since the martingale problem with operator \( L^1 N \) has a unique solution, the associated differential equation has a unique solution for each initial condition and \( \alpha^0(\cdot) \) is unique. For each \( T < \infty \) and \( t \leq T \), \( P^0(\cdot) \) agrees with \( P^N(\cdot) \) on all Borel subsets of the set of paths in \( D(0, \infty) \) with values in \( S_N \). By using (13)–(21), \( \alpha^N(\cdot) \) to \( \psi^N(\cdot) \), we have \( \psi^\mu(\cdot) \) converges weakly to \( \psi(\cdot) \). The proof of Theorem III.1 is completed.

\[\square\]

**Remark III.4** The following calculation will be used for both the slow and fast Markov chain cases. The result is essentially one about two-time-scale Markov chains considered in [27]. Define a probability vector by \( p^\varepsilon(t) = (p(\alpha_n = \alpha_1 \ldots, p(\alpha_n = \alpha_m)) \in \mathbb{R}^{1 \times m_0} \). Note that \( p^\varepsilon_0 = (p_0, \ldots, p_0, m_0) \). Because the Markov chain is time homogeneous, \( (P^\varepsilon)^N \) is the asymptotic transition probability matrix with \( P^\varepsilon = I + \varepsilon Q \). The \( 1 \times m_0 \) dimensional probability vector \( p(t) = (p_1(t), \ldots, p_{m_0}(t)) \) of the continuous Markov chain, for \( t \geq 0 \), satisfies the Chapman-Kolmogorov equation
\[
d p(t) \frac{dt}{dt} = p(t) Q, \quad p(0) = p_0, \tag{22}
\]
where \( p_0 \) is the initial probability. Then, for some \( 0 < \lambda_1 < 1 \), \( p^\varepsilon(t) = p(\varepsilon n) + O(\varepsilon + \lambda_1^{-n}) \), \( 0 \leq n \leq O(1/\varepsilon) \), where \( p(\varepsilon n) \) is defined in (22). In addition, \( (P^\varepsilon)^{n_0} = \Xi(\varepsilon n_0, \varepsilon n) + O(\varepsilon + \lambda_1^{-n_0}) \), where \( t_0 = \varepsilon n_0 \) and \( t = \varepsilon n \), \( \Xi(t_0, t) \) satisfies
\[
\begin{cases}
d \Xi(t_0, t) \frac{dt}{dt} = \Xi(t_0, t) Q, \\
\Xi(t_0, t_0) = 1. \tag{23}
\end{cases}
\]
Define the continuous-time interpolation \( \alpha^\varepsilon(t) \) of \( \alpha_n \) as \( \alpha^\varepsilon(t) = \alpha_n \) for \( t \in [n\varepsilon, n\varepsilon + \varepsilon) \). Then \( \alpha^\varepsilon(\cdot) \) converges weakly to \( \alpha(\cdot) \), which is a continuous-time Markov chain generated by \( Q \) with state space \( M \). The \( E\alpha_n \) can be approximated by: \( E\alpha_n = \pi(\varepsilon n) + O(\varepsilon + \lambda_1^{-n}) \), for \( n \leq O(1/\varepsilon) \), \( \pi(\varepsilon n) \equiv \sum_{j=1}^{m_0} a_j p_j(\varepsilon n) \).

### 3.2. Slowly-Varying Markov Chain: \( \varepsilon \ll \mu \)

The rationale here is as follows. Since the Markov chain changes so slowly, the time-varying parameter process is essentially a constant. To facilitate the discussion and to fix notation, we take \( \varepsilon = \mu^{1+\Delta} \) for some \( \Delta > 0 \) in what follows.

To analyze the algorithm, we define the continuous-time interpolations the same as before. We still need to use the \( N \)-truncation device as in the previous section. However, for notational simplicity, we assume that all the iterates are bounded. The tightness of \( (\psi^\mu, \alpha^\mu(\cdot)) \) can be proved as in the previous case. To figure out the limit, while the estimates of the other terms are the same as in the previous section, we need only treat the average of the term involving the Markov chain \( \alpha_k \). In fact,
\[
\sum_{l_{\mu}=t}^{t+s} \frac{1}{m_{\mu}} \sum_{k=l_{\mu}}^{l_{\mu}+m_{\mu}-1} \text{Sgn}(\varphi_k) \varphi_k' \alpha_k 
= \sum_{l_{\mu}=t}^{t+s} \frac{1}{m_{\mu}} \sum_{k=l_{\mu}}^{l_{\mu}+m_{\mu}-1} (\text{Sgn}(\varphi_k) \varphi_k' - G) \alpha_k 
+ \sum_{l_{\mu}=t}^{t+s} \frac{1}{m_{\mu}} \sum_{k=l_{\mu}}^{l_{\mu}+m_{\mu}-1} G \alpha_k. \tag{24}
\]
Thus we can show as in the previous section that the limit of the term on the first line of (24) is the same as that of the last term on the second line. We therefore
concentrate on the last term of (24), which is equal to
\[
Eh_N^\mu \sum_{m=1}^{m_0} \sum_{\ell_m=t}^{t+s} \delta_m \nabla \ell (\varsigma_{m_\mu}, \alpha_{l_\mu}) \\
\times \left( \frac{1}{m_\mu} \sum_{k=1}^{l_\mu} G a_j \epsilon_k = a_j | \alpha_{l_\mu} = a_{i_1} \right) \\
\times P(\alpha_{l_\mu} = a_i | \alpha_0 = a_{i_0}) p^0 \\
\to Eh_N \sum_{i=1}^{m_0} \int_{t+s}^{t+\epsilon} \nabla \ell (\varsigma(\tau), \alpha(\tau)) G a_i \epsilon k \epsilon_l \epsilon_m d \tau,
\]
where \( p^0 = P(\alpha_0 = a_{i_0}) \). In the above, we have used Remark III.4 and \( \epsilon = \mu^{1+\Delta} \) for some \( \Delta > 0 \) to obtain that for some \( 0 < \lambda_1 < 1 \),
\[
(\mathbb{P} \epsilon)^{k-l_{\mu}} = \Xi(\epsilon k, \epsilon l_{\mu}) + O(\epsilon + \lambda_1^{-1}(k-l_{\mu})) \\
\to I \text{ as } \mu \to 0,
\]
Likewise, \((\mathbb{P} \epsilon)^{l_{\mu}} \to I \text{ as } \mu \to 0\). We omit the details, but present the main result as follows,

**Theorem III.5** Assume the conditions of Theorem III.1 are satisfied with the modification \( \epsilon = \mu^{1+\Delta} \) for some \( \Delta > 0 \). Then we have \( \varsigma^\mu(\cdot) \) converges weakly to \( \varsigma(\cdot) \) such that \( \varsigma(\cdot) \) is the unique solution of the differential equation
\[
\frac{d}{dt} \varsigma(t) = G(\varsigma(t) - \sum_{i=1}^{m_0} a_i P(\alpha_0 = a_i)), \quad \varsigma(0) = \varsigma_0.
\]

We further obtain the following asymptotic result,

**Corollary III.6** Under the conditions of Theorem III.5, assume further \( G \) is a stable matrix and \( \{\varsigma_n\} \) is tight. Then for any \( t_\mu \to \infty \) as \( \mu \to 0 \), \( \varsigma^\mu(\cdot + t_\mu) \) converges weakly to \( \varsigma_\ast = \sum_{i=1}^{m_0} a_i P(\alpha_0 = a_i) \).

**Proof.** For any \( T < \infty \), consider the pair \( (\varsigma^\mu(\cdot + t_\mu), \varsigma^\mu(\cdot + t_\mu - T)) \). As previously, it can be shown that the sequence is tight. We can then extract convergent subsequences. Take such as sequence and denote the limit by \( (\varsigma(\cdot), \varsigma(\cdot)) \). Note \( \varsigma(0) = \varsigma(0) \). The value of \( \varsigma(0) \) may not be known, but the set of possible \( \{\varsigma(0)\} \) is tight since \( \{\varsigma_n\} \) is tight. Consequently, we have
\[
\varsigma(T) = \exp(GT) \varsigma(0) - \int_0^T \exp(G(T - s)) \alpha ds.
\]
Making a change of variable \( t = T - s \) in the right-hand side above, we arrive at
\[
\varsigma(T) = \exp(GT) \varsigma(0) + \int_0^T \exp(Gt) \alpha dt \\
\to \varsigma_\ast \text{ as } T \to \infty.
\]

The desired result thus follows.

**Remark III.7** In the above theorem, we have assumed the tightness of \( \{\varsigma_n\} \) in \( \mathbb{R}^n \), which can be proved. In Section IV, we prove a mean squares estimate. The techniques used there can be adopted to obtain the tightness.

### 3.3. Fast-Varying Markov Chain: \( \mu \ll \epsilon \)

The idea for the fast varying chain is that the parameter changes so fast that it comes quickly to the stationary distribution of the Markov chain. As a result, the limit dynamic system is one that is averaged out with respect to the stationary distribution of the Markov chain. In this section, we take \( \epsilon = \mu^\gamma \) where \( 1/2 < \gamma < 1 \). Then, letting \( \mu l_{\mu} \to \tau \) as in the proof of Theorem III.1, we have \( \epsilon(k-l_{\mu}) = \mu^\gamma(k-l_{\mu}) \to \infty \). Thus, for some \( 0 < \lambda_1 < 1 \), \( \Xi_{ij}(\epsilon l_{\mu}, \epsilon k) = \nu_j + O(\epsilon + \lambda_1^{-1}(k-l_{\mu})) \), where \( \nu = (\nu_1, \ldots, \nu_{m_0}) \) is the stationary distribution of the continuous-time Markov chain with generator \( Q \). \( \Xi_{ij}(s_1, s_2) \) denotes the \( ij \)th entry of the matrix \( \Xi(s_1, s_2) \). Therefore, we can show that as \( \mu \to 0 \),
\[
Eh_N^\mu \sum_{i=1}^{m_0} \sum_{\ell_m=t}^{t+s} \delta_m \nabla \ell (\varsigma_{m_\mu}, \alpha_{l_\mu}) \\
\times \left( \frac{1}{m_\mu} \sum_{k=1}^{l_\mu} G a_j \epsilon_k = a_j | \alpha_{l_\mu} = a_{i_1} \right) \\
\times P(\alpha_{l_\mu} = a_i | \alpha_0 = a_{i_0}) p^0 \\
\to Eh_N (\varsigma(\tau), \alpha(\tau)) G(\varsigma(t) - \sum_{i=1}^{m_0} a_i \nu_i). \tag{28}
\]

As a result, the following holds,

**Theorem III.8** Assume the conditions of Theorem III.1 are satisfied with the modification of \( \epsilon = \mu^\gamma \) for some \( 1/2 < \gamma < 1 \). Then we have \( \varsigma^\mu(\cdot) \) converges weakly to \( \varsigma(\cdot) \) such that \( \varsigma(\cdot) \) is the unique solution of the differential equation
\[
\frac{d}{dt} \varsigma(t) = G(\varsigma(t) - \sum_{i=1}^{m_0} \nu_i a_i), \quad \varsigma(0) = \varsigma_0. \tag{29}
\]

Similar to the slow chain case, we obtain the following corollary, whose proof is omitted.

**Corollary III.9** Under the conditions of Theorem III.8, assume further \( G \) is a stable matrix and \( \{\varsigma_n\} \) is tight for sufficiently large \( n \). Then for any \( t_\mu \to \infty \) as \( \mu \to 0 \), \( \varsigma^\mu(\cdot + t_\mu) \) converges weakly to \( \bar{\alpha} = \sum_{i=1}^{m_0} \nu_i a_i \).
IV. Mean Squares Order Estimates

Define the sequence $\zeta_n = \xi_n - \alpha_n$. This section is devoted to the mean squares estimate of $\zeta_n$. It establishes the tracking ability with the precise order estimate. Note that we do not have restriction of the relative sizes of $\varepsilon$ and $\mu$ as far as the statement of the theorem. Nevertheless, from the mean-squares type estimate, the relative size of $\mu$ and $\varepsilon$ needs to satisfy certain conditions. We will come back to this point after the result is established.

**Theorem IV.1** Assume (A1) and (A2). Then for sufficiently large $n$,

$$E|\zeta_n|^2 = E|\xi_n - \alpha_n|^2 = O\left(\mu + \varepsilon + \varepsilon^2/\mu\right).$$

(30)

**Remark IV.2** In the statement of Theorem IV.1, by “for sufficiently large $n$, (30) holds” we mean that there is an $N_\mu$ such that for all $n \geq N_\mu$, (30) holds. This convention will be used in what follows.

**Proof of Theorem IV.1.** We begin by defining a Liapunov function $V(x) = (x'x)/2$. Use $E_n$ to denote the conditional expectation with respect to the $\sigma$-algebra $\mathcal{F}_n$. Then

$$E_nV(\zeta_{n+1}) - V(\xi_n) = E_n[\zeta_n'|-\mu \text{Sgn}(\phi_n)\phi_n'\zeta_n + \mu \text{Sgn}(\phi_n)e_n + (\alpha_n - \alpha_{n+1})^2].$$

(31)

In view of the Markovian assumption, the independence of the Markov chain with the signals $\{(\phi_n, e_n)\}$, and the structure of the transition probability matrix (4),

$$E_n(\alpha_n - \alpha_{n+1}) = \sum_{i=1}^{n} \left[ a_i - \sum_{j=1}^{n} a_j (\delta_{ij} + \varepsilon q_{ij}) \right] I_{\alpha_n = a_i} = O(\varepsilon).$$

(32)

Note that we are taking conditional expectation w.r.t. $\mathcal{F}_n$, and $I_{\alpha_n = a_i}$ is $\mathcal{F}_n$-measurable. Similarly, we also have

$$E_n|\alpha_n - \alpha_{n+1}|^2 = O(\varepsilon).$$

(33)

Using an elementary inequality $ab \leq (a^2 + b^2)/2$ for two real numbers $a$ and $b$, we have $|\zeta_n| = |\zeta_n| \cdot 1 \leq (|\zeta_n|^2 + 1)/2$, so

$$O(\varepsilon)|\zeta_n| \leq O(\varepsilon)(V(\zeta_n) + 1).$$

(34)

The boundedness of the signal $\{(\phi_n, e_n)\}$ then yields

$$E_n| - \mu \text{Sgn}(\phi_n)\phi_n'\zeta_n + \mu \text{Sgn}(\phi_n)e_n + (\alpha_n - \alpha_{n+1})|^2$$

$$= E_n|\alpha_n - \alpha_{n+1}|^2 + O(\mu^2 + \mu \varepsilon)(V(\zeta_n) + 1).$$

(35)

Using (33), (34), and (35), we obtain

$$E_nV(\zeta_{n+1}) - V(\xi_n) = E_n[\zeta_n'|-\mu \text{Sgn}(\phi_n)\phi_n'\zeta_n + \mu \text{Sgn}(\phi_n)e_n + (\alpha_n - \alpha_{n+1})^2] + O(\mu^2 + \mu \varepsilon)(V(\zeta_n) + 1).$$

(36)

We proceed to treat the next to the last line and the first term on the last line of (36) by using perturbed Liapunov functions [14]. Define three perturbations of the Liapunov function by

$$V_1^\mu(\zeta, n) = -\mu \sum_{j=n}^{\infty} E_n \tilde{\zeta}(\text{Sgn}(\phi_j) \phi_j' - G)\tilde{\zeta},$$

$$V_2^\mu(\zeta, n) = \mu \sum_{j=n}^{\infty} E_n \tilde{\zeta}(\text{Sgn}(\phi_j)e_j),$$

$$V_3^\mu(\zeta, n) = \sum_{j=n}^{\infty} E_n(\alpha_j - \alpha_{j+1}),$$

$$V_4^\mu(n) = \sum_{j=n}^{\infty} E_n(\alpha_n - \alpha_{n+1})'(\alpha_j - \alpha_{j+1}).$$

For each $\zeta$, by virtue of (A2), it is easily verified that

$$\mu \sum_{j=n}^{\infty} E_n|\text{Sgn}(\phi_j) \phi_j' - G| \tilde{\zeta}^2 \leq O(\mu)(V(\zeta) + 1),$$

so

$$|V_1^\mu(\zeta, n)| \leq O(\mu)(V(\zeta) + 1).$$

(38)

Similarly, for each $\zeta$,

$$|V_2^\mu(\zeta, n)| \leq O(\mu)(V(\zeta) + 1).$$

(39)

Note the irreducibility of $Q$ implies that of $I + \varepsilon Q$. It can be shown that $\sum_{j=n}^{\infty} |(I + \varepsilon Q)^{j+1-n} - (I + \varepsilon Q)^{j-n}| = O(\varepsilon).$ Thus

$$|V_3^\mu(\zeta, n)| \leq O(\varepsilon)(V(\zeta) + 1), \quad |V_4^\mu(n)| = O(\varepsilon).$$

(40)

Note also that

$$E_nV_1^\mu(\zeta_{n+1}, n + 1) - V_1^\mu(\zeta_n, n) = E_nV_1^\mu(\zeta_{n+1}, n + 1) - E_nV_1^\mu(\zeta_n, n + 1) + E_nV_1^\mu(\zeta_n, n + 1) - V_1^\mu(\zeta_n, n).$$

(41)

It follows that

$$E_nV_1^\mu(\zeta_{n+1}, n + 1) - V_1^\mu(\zeta_n, n) = \mu E_n\tilde{\zeta}(\text{Sgn}(\phi_j) \phi_j' - G)\tilde{\zeta},$$

(42)
Similarly, we obtain
\[ E_n V^\mu((\tilde{z}_{n+1}, n + 1) - E_n V^\mu((\tilde{z}_n, n + 1)) \]
\[ = -\mu \sum_{j=n+1}^{\infty} E_n [\tilde{z}_{n+1} - \tilde{z}_n] [E_n V_{\phi_j}^\mu g_j] \]
\[ = -\mu \sum_{j=n+1}^{\infty} E_n [\tilde{z}_{n+1} - \tilde{z}_n] [E_{\phi_j} V_n^\mu g_n] + O(\epsilon). \]

Moreover,
\[ |\mu \sum_{j=n+1}^{\infty} E_n [\tilde{z}_{n+1} - \tilde{z}_n] [E_n V_{\phi_j}^\mu g_n] | \]
\[ \leq O(\mu^2 + \mu)(V(\tilde{z}_n) + 1), \]
\[ |\mu \sum_{j=n+1}^{\infty} E_n [\tilde{z}_{n+1} - \tilde{z}_n] [E_n V_{\phi_j}^\mu g_n] | \]
\[ \leq O(\mu^2 + \mu)(V(\tilde{z}_n) + 1). \]

Therefore, we obtain
\[ E_n V^\mu((\tilde{z}_{n+1}, n + 1) - V^\mu((\tilde{z}_n, n)) \]
\[ = \mu E_n [E_n V_{\phi_j}^\mu g_n] - G(\tilde{z}_{n+1} - \tilde{z}_n) \]
\[ + O(\mu^2 + \mu)(V(\tilde{z}_n) + 1). \]

Similarly, we obtain
\[ E_n V^\mu((\tilde{z}_{n+1}, n + 1) - V^\mu((\tilde{z}_n, n)) \]
\[ = -\mu E_n [E_n V_{\phi_j}^\mu g_n] + O(\mu^2 + \mu)(V(\tilde{z}_n) + 1), \]
\[ E_n [V^\mu((\tilde{z}_{n+1}, n + 1) - V^\mu((\tilde{z}_n, n + 1))] \]
\[ = O(\epsilon^2 + \mu^2)(V(\tilde{z}_n) + 1), \]
\[ E_n V^\mu((\tilde{z}_n, n + 1) - V^\mu((\tilde{z}_n, n)) \]
\[ = -E_n [E_n V_{\phi_j}^\mu g_n] - (\alpha_n + \alpha_{n+1}), \]
\[ E_n V^\mu((\tilde{z}_n, n + 1)) - V^\mu((\tilde{z}_n, n)) \]
\[ = O(\epsilon^2) - E_n [E_n V_{\phi_j}^\mu g_n] = -E_n [E_n V_{\phi_j}^\mu g_n] - (\alpha_n + \alpha_{n+1}). \]

Define
\[ W(\tilde{z}, n) = V(\tilde{z}) + \sum_{i=1}^{3} V_i^\mu((\tilde{z}, n)) + V_i^\mu((\tilde{z}, n)). \]

Since \( G \) is a stable matrix, there is a \( \lambda > 0 \) such that \( \tilde{\xi} G \tilde{\xi} \geq \lambda V(\tilde{\xi}) \). Then using (36) and (46)–(47), upon cancelation, we obtain
\[ E_n W((\tilde{z}_{n+1}, n + 1) - W((\tilde{z}_n, n)) \]
\[ \leq -\lambda \mu V((\tilde{z}_n) + O(\mu^2 + \epsilon^2)(V(\tilde{z}_n) + 1) \]
\[ \leq -\lambda \mu W((\tilde{z}_n, n) + O(\mu^2 + \epsilon^2)(W(\tilde{z}_n, n) + 1). \]

In (48), from the third line to the last line, we used estimates (38)–(40) and replaced \( V(\tilde{z}_n) \) by \( W(\tilde{z}_n, n) \), which results in an \( O(\mu \epsilon) \) term by the boundedness of \( \tilde{\xi} \); we also used \( O(\mu \epsilon) = O(\mu^2 + \epsilon^2) \) via an elementary inequality.

Choose \( \mu \) and \( \epsilon \) small enough so that there is an \( \lambda_0 > 0 \) satisfying \( \lambda_0 \leq \lambda \) and \( -\lambda \mu + O(\mu^2) + O(\epsilon^2) \leq -\lambda_0 \mu \). Then we obtain
\[ E_n W((\tilde{z}_{n+1}, n + 1) \]
\[ \leq (1 - \lambda_0 \mu) W((\tilde{z}_n, n) + O(\mu^2 + \epsilon^2)). \]

Taking expectation above and iterating on the resulting inequality yield
\[ EW((\tilde{z}_{n+1}, n + 1) \]
\[ \leq (1 - \lambda_0 \mu)(n - N_0) W((\tilde{z}_N, n) + O(\mu^2 + \epsilon^2)). \]

By taking \( n \) large enough, we can make \( (1 - \lambda_0 \mu)(n - N_0) \leq O(\mu \epsilon) \). Thus \( EW((\tilde{z}_{n+1}, n + 1) \leq O(\mu + \epsilon^2 / \mu) \). Finally, applying (38)–(40) again, we also obtain \( EW((\tilde{z}_{n+1}) \leq O(\mu + \epsilon^2 / \mu) \). Thus the desired result follows.

\[ \square \]

**Corollary IV.3** Under the conditions of Theorem IV.1, we have the following results: For sufficiently large \( n \) (i.e., there is an \( N_\mu \) such that for all \( n \geq N_\mu \),

(i) When \( \epsilon = O(\mu) \), \( E[\tilde{z}_n - \alpha_n]^2 \leq O(\mu) = O(\epsilon). \)

(ii) When \( \epsilon = O(\mu^{1+\Delta}) \) for some \( 0 < \Delta \leq 1 \), \( E[\tilde{z}_n - \alpha_n]^2 \leq O(\mu^{1/2+\frac{\Delta}{2}}). \)

(iii) When \( \epsilon = O(\mu^\gamma) \) for some \( 0 < \gamma < 1 \), \( E[\tilde{z}_n - \alpha_n]^2 \leq O(\mu^{\gamma} \land \mu^{2\gamma - 1}) \), where \( a \land b = \min(a, b) \) for \( a, b \in \mathbb{R} \).

**V. Asymptotic Distributions**

This section again treats three parts in accordance with the relative orders of \( \epsilon \) and \( \mu \). As a standing condition, we assume the following condition holds throughout this section.

(A3) Assume that \( \sqrt{\mu} \sum_{j=1}^{(t/\mu) - 1} Sgn(\phi_j) e_j \) converges weakly to a Brownian motion \( \tilde{w}(t) \) with covariance \( \tilde{\Sigma} t \).

Note that in the above, we have assumed the weak convergence to a Brownian motion process. Suppose that, for example, \( Sgn(\phi_n) e_n \) is a stationary mixing process satisfying \( \sum_{j=1}^{\infty} \phi_j^{1/2} < \infty \) where \( \phi_n \) is the associated mixing measure. Then we can in fact derive \( \sqrt{\mu} \sum_{j=1}^{(t/\mu) - 1} \omega_j \) converges weakly to a Brownian motion \( \tilde{w}(t) \) with covariance \( \tilde{\Sigma} t \) such that the covariance \( \tilde{\Sigma} \) is given by \( \tilde{\Sigma} = E \omega_0 \omega'_0 + \sum_{j=1}^{\infty} E \omega_j \omega'_0 + \sum_{j=1}^{\infty} E \omega_0 \omega'_j \). We refer the reader to [3] for further details.
5.1. Scaled Errors: $\varepsilon = \mu$

By virtue of Corollary IV.3 and Remark IV.2, there is an $N_\varepsilon = N_\mu$ such that for all $n \geq N_\mu, E|\varepsilon_n - \alpha_n|^2 = O(\mu) = O(\varepsilon)$. To investigate the rate of variation of the sequence $\varepsilon_n - \alpha_n$, we consider $u_n = (\varepsilon_n - \alpha_n)/\sqrt{\mu}$. Using the above result, we can in fact establish that $\{u_n : n \geq N_\mu\}$ is tight. To proceed, define $w^{(t)} = u_n$ for $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$. We then can proceed to the study of the asymptotic distribution of $w^{(t)}$. A natural thing to do is again to take an $N$-truncation as done in Section III. The steps will be similar to the previous section. For brevity, we omit the verbatim proof but state the following theorem.

**Theorem V.1** Assume the conditions of Theorem III.1 and (A3) hold. Then $w^{(t)}$ converges weakly to $u^{(t)}$ such that $u^{(t)}$ is the solution of

$$du(t) = G\zeta dt + \tilde{\Sigma}^{1/2}dw,$$

where $w^{(t)}$ is a standard $\mathbb{R}^r$-valued Brownian motion.

We note that (49) has a unique solution to each initial condition since it is linear in $u$. To obtain the weak convergence, we show that $u^{(t)}$ is the solution of the martingale problem with operator

$$\mathcal{L}f(u) = \nabla f(u)G\zeta u + \frac{1}{2}tr(\tilde{\Sigma}\nabla^2 f(u)),$$

where $\nabla^2 f(u)$ denotes the Hessian of $f$. The rest of the detailed developments are similar to that of the proof of Theorem III.1; we omit the details.

5.2. Scaled Errors: $\varepsilon \ll \mu$ and $\mu \ll \varepsilon$

Here we treat the remaining two cases. The developments are similar to the previous. We shall be brief with the concentration on the main results.

**Slowly-Varying Markov Chain Case:** $\varepsilon \ll \mu$ Take $\varepsilon = \mu^{1+\Delta}$ as before for definiteness. Define $v_n = (\varepsilon_n - \alpha_n)/\sqrt{\mu}$, $v^{(t)} = v_n$ for $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$. We obtain the following result.

**Theorem V.2** Assume the conditions of Theorem III.5 and (A3) hold. Then $v^{(t)}$ converges weakly to $v^{(t)}$ such that $v^{(t)}$ is the solution of

$$dv(t) = G\zeta dt + \tilde{\Sigma}^{1/2}dw,$$

where $v^{(t)}$ is a standard $\mathbb{R}^r$-valued Brownian motion.

**Fast-Varying Markov Chain Case:** $\mu \ll \varepsilon$ As before, we take $\varepsilon = \mu^\gamma$ for some $1/2 < \gamma < 1$. Define $z_n = (\varepsilon_n - \overline{\alpha})/\sqrt{\mu}$ and $z^{(t)} = v_n$ for $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$. We obtain the following result.

**Theorem V.3** Assume the conditions of Theorem III.8 and (A3) hold. Then $z^{(t)}$ converges weakly to $v^{(t)}$ such that $z^{(t)}$ is the solution of

$$dz(t) = G\zeta dt + \tilde{\Sigma}^{1/2}dw,$$

where $w^{(t)}$ is a standard $\mathbb{R}^r$-valued Brownian motion.

**Remark V.4** Let us give a brief account on the three results Theorem V.1, Theorem V.2, and Theorem V.3. The three limit stochastic differential equations all have the same form. The stochastic parts are the same in each case. However, they have very different meanings. Theorem V.1 is about tracking error of the approximation sequence to that of the Markov chain. In this case, the adaptation rate and the transition frequency of the Markov chain are the same. In Theorem V.2, we examine the asymptotics of $\varepsilon_n - \alpha_n$. Roughly, it indicates that $\varepsilon_n - \alpha_n$ is asymptotically normal with mean 0 and covariance $\mu S$, where $S$ is the solution of the Lapunov equation $GS + SG^* = -\Sigma$. Alternatively, this asymptotic covariance may be represented by $S = \int_0^\infty \exp(Gs)\Sigma \exp(G^*s)ds$. Likewise, Theorem V.3 is with regard to the fast changing Markov chain case. The asymptotic properties of interest here is $\varepsilon_n - \overline{\alpha}$.

6. Numerical Experiments and Further Remarks

6.1. Simulation Study

This section demonstrates the tracking property of the algorithm via simulations of each of the three cases. We fix the step size $\mu = 0.1$ and take $\varepsilon = \mu^{1+\Delta}$ ($\varepsilon = O(\mu)$), $\varepsilon = \mu^\gamma$ (Slowly-Varying Markov Chain), and $\varepsilon = \sqrt{\mu}$ (Fast-Varying Markov Chain), respectively. We let the state space for the parameter $\alpha_n$ be $\mathcal{M} = \{-1, 0, 1\}$, and the transition matrix be (4) with the generator of the continuous-time Markov chain given by $Q = \begin{bmatrix} -0.6 & 0.4 & 0.2 \\ 0.2 & -0.5 & 0.3 \\ 0.4 & 0.1 & -0.5 \end{bmatrix}$. Associated with the continuous-time Markov chain whose generator is $Q$, the stationary distribution is a uniform distribution $\nu = (1/3, 1/3, 1/3)$. Corresponding to this case, $\pi = \sum_{i=1}^{3} a_i \nu_i = 0$. We take the initial distribution for $\alpha_0$...
to be \((3/4, 1/8, 1/8)\) which gives \(\alpha_* = \sum_{i=1}^{3} a_i P(\alpha_0 = a_i) = -0.625\). We take \(\{\varphi_n\}\) as \(\mathcal{N}(0, 1)\) and \(\{e_n\}\) as \(\mathcal{N}(0, 0.4)\) (both i.i.d). We observe 1,000 iterations of the parameter \(\alpha_n\) and the estimate \(\varsigma_n\) for cases \(\varepsilon = O(\mu)\) and \(\varepsilon \gg \mu\), and we observe 10,000 iterations for the case \(\varepsilon \ll \mu\) in order to see some variation.

Adaptive using sign regressor performs remarkably well. When the adaptation rate is in line with the transition frequency \(\varepsilon = O(\mu)\), our algorithm still shows good tracking properties. When we have fast changing Markov chain, the graph shows that it is impossible to track the variations. In this case, a sensible thing to do is to examine the average behavior. As demonstrated in the previous section, our estimate is in fact an average with respect to the stationary distribution of the Markov chain associated with the generator \(Q\). Finally, we present a picture of the scale estimation error in Figure 4, in which the diffusion behavior is clearly pronounced.

6.2. Further Remarks

This paper presents asymptotic properties of adaptive filtering algorithms with sign regressors for tracking randomly varying system parameters. Different combinations of stepsizes of adaptation and transition rates of the time-varying random process create three distinct scenarios. Our results reveal the basic properties of the associated stochastic systems and the capability and limitations of the tracking algorithms. Adaptive filtering and stochastic recursive algorithms have been used widely for a wide variety of applications in adaptive signal processing and adaptive control. New areas and related applications also include iterative learning [20], system identification with quantized observations [22], and networked systems among others. To further our understanding and to enlarge the applicability, in the subsequent studies, we aim to study sign-error and sign-sign algorithms as well.

REFERENCES

1. A. Benveniste, M. Gouoursat, and G. Ruget, Analysis of stochastic approximation schemes with discontinuous and dependent forcing terms with...


