DIGITALCOMMONS
—@WAYNESTATE—

Wayne State University

Wayne State University Dissertations

1-1-2014

Adaptive Stochastic Systems: Estimation, Filtering, And Noise Attenuation

Araz Ryan Hashemi
Wayne State University,

Follow this and additional works at: http://digitalcommons.wayne.edu/oa_dissertations

Part of the Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

ADAPTIVE STOCHASTIC SYSTEMS:
ESTIMATION, FILTERING, AND NOISE ATTENUATION

by

ARAZ HASHEMI

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

2014

MAJOR: MATHEMATICS

Approved by:

_____

Advisor   Date

_____

_____

_____

DEDICATION

To my parents

ACKNOWLEDGEMENTS

It would be impossible to adequately express my gratitude and appreciation for my advisor, Professor George Yin. His ceaseless devotion, support, and encouragement has been the driving force behind all my accomplishments during my doctoral studies. He is the inspiration for the kind of mathematician I aspire to be.

I must also thank Professors Le Yi Wang, Kazuhiko Shinki, and Tze-Chien Sun for serving on my committee. In addition, they have all been instrumental in introducing me to the topics pursued in this work.

Of course, nothing would have been possible without the enduring support (and patience) of my family. Their love and encouragement has pulled me through even the most trying challenges I encountered during my studies.

I have been with the Wayne State University Mathematics Department for many years, and the relationships I formed here have developed me into the mathematician and indeed the man I have become. I could fill volumes with my appreciation of the many faculty, staff, and students which have been like family to me. Above all else, I must express my gratitude for the support and friendship of the late James Veneri. I often wonder at how different my life would be if Jim hadn't taken me into his carefully-cultivated community. He gave me direction when I was at a crossroad in my life, and without his help I shouldn't have come half as far as I have. He was a remarkable man and is greatly missed.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

This dissertation investigates problems arising in identification and control of stochastic systems, in which random noise corrupts the observations of the system. The focus is on developing methods that are *adaptive* in nature; that is, methods which can respond to changes in the dynamics of the underlying systems. Adaptive filtering algorithms use feedback (usually in the form of error) to iteratively adjust its estimates of the system parameters. Because of their recursive form and ability to track time-varying parameters, adaptive filters have been an important tool in many recent technologies and applications. Examples include tuning of manufacturing systems, navigation and target tracking in autonomous vehicles, financial modeling across switching market dynamics, and data shuffling in communication networks. Adaptive filtering has been especially effective in CDMA (code-division multiple access) wireless communication networks [17] for filtering a given user's signal from the other signals being concurrently transmitted as well as ambient noise across the communication channel.

We begin by considering linear systems whose coefficients evolve as a slowly-varying Markov Chain. These slow Markov models are useful for modeling systems whose dynamics change infrequently (in relation to the signal/sampling rate), yet whose parameters 'jump' large distances whenever a transition occurs. Again, communication networks are a natural candidate for such models because of switching network topologies resulting from channel connections, signal interruptions, transmission queueing and routing dynamics.

We analyze families of constant step-size (or gain size) algorithms for estimating and tracking the coefficient parameter in the Markovian setting: the Least-Mean Squares (LMS), Sign-Regressor (SR), and Sign-Error (SE) algorithms. While the LMS algorithm was studied in [29], we consider analysis of its (faster) variants the SR algorithm in Chapter 2 and the SE algorithm in Chapter 3. The analysis is carried out in a multi-scale framework considering the relative size of the gain (rate of adaptation) to the transition rate of the Markovian system parameter. Mean-square error bounds are established, and weak convergence methods are

employed to show the convergence of suitably interpolated sequences of estimates to solutions of systems of ordinary and stochastic differential equations with regime switching. Simulation studies are presented to display the tracking properties corresponding to the relationship between the adaptation rate of the algorithm and the transition rate of the underlying Markov chain.

Next, in Chapter 4 we consider problems in noise attenuation in systems with unmodeled dynamics and stochastic signal measurement errors. Unmodeled dynamics must be considered when the modeled system order does not account for the full system dynamics. A robust two-phase design procedure of the stochastic approximation type is developed which first estimates the signal in a simplified form, and then applies a control to tune out the noise. Worst-case error bounds are derived in terms of the unmodeled dynamics and variances of the disturbance and measurement errors. Simulation studies are then given to display the noise attenuation performance of the algorithm.

Finally, in Chapter 5, we summarize the theme of this work with some further remarks and present some directions for future work.

2　Sign-Regressor Algorithms for Markovian Parameters

2.1　Motivation and Formulation

Consider the multiple input, single output adaptive filtering problem for the system of signals given by

$$y_n = \varphi_n' \alpha_n + e_n, \qquad n \in \mathbb{N} \tag{2.1}$$

where $\varphi_n \in \mathbb{R}^r$ is the sequence of input regression vectors (possibly stochastic), $y_n \in \mathbb{R}$ are the corresponding observation signals, $e_n \in \mathbb{R}$ is a sequence of zero-mean error signals (noise), and α_n is the time-varying parameter process.

Linear systems with a constant parameter $\alpha_n \equiv \alpha^*$ are very well known in classical statistics and signal processing, and copious amounts of results are available for efficient estimation and identification. Time-varying systems such as (2.1) have also been extensively studied (see [4, 12, 19, 22]), but the usual approach assumes the parameter process evolves either deterministic continuously or stochastically due to some zero mean Guassian disturbance. These models assume that parameter changes are small when they occur, which allows for more tractability in establishing convergence or error bounds.

In contrast, we analyze the behavior of systems where the parameter process α_n acts as a "slow" Markov chain, meaning that the parameter randomly "jumps" large distances between many possible states in the state space (albeit infrequently). More precisely, the Markov chain α_n has a near-identity transition matrix $P^\varepsilon = I + \varepsilon Q$ for some jump frequency parameter ε and a matrix Q which is a generator of a continuous-time Markov chain. The smaller ε is, the closer the transition matrix P^ε is to the identity matrix I, implying α_n jumps between states less frequently. Conversely, the larger the value of ε is the more frequently α_n can jump. Hence the parameter ε shall be referred to as the *transition rate* of the Markov chain α_n.

Many stochastic systems have randomly time-varying parameters can be best described by this slow Markov chain model. For example, networked systems include communication channels as part of the system topology. Channel connections, interruptions, data transmission queuing and routing, packet delays and losses, are always random. Markov chain models become a natural choice for such systems. In [29] the problem of the adaptive multiuser detector is considered for a synchronous CDMA-DS system with a maximum of N users. The optimal multiuser detector is dependent on the current active user set and hence can be modeled as a Markov chain with 2^{N-1} states; see [27, 34] for further examples of Markov

chain system models. For control strategy adaptation and performance optimization, it is essential to capture time-varying system parameters during their operations, which leads to the problems of identifying Markovian regime-switching systems pursued here.

For the adaptive filtering problem, the goal is to use known input values of φ_n (e.g. from a training sequence) and observed output values y_n to estimate and track the underlying system parameter α_n. Stochastic approximation algorithms of the Robbins-Monro type [21] have been widely used to generate recursive estimates θ_n for systems such as (2.1). A traditional RM algorithm known as the Least Mean Squares algorithm minimizes the expected norm-squared error between the actual and predicted signals $\mathbb{E}|y_n - \varphi_n'\theta_n|^2$ is given as follows.

Algorithm 1 (Least Mean Squares). *The Least Mean Squares (LMS) algorithm for the adaptive filtering problem given by (2.1) recursively generates estimates θ_n of α_n by*

$$\theta_{n+1} = \theta_n + \mu\varphi_n\left(y_n - \varphi_n'\theta_n\right) \tag{2.2}$$

The parameter μ in (2.2) is the step-size (gain) of the the algorithm which controls the magnitude of the change between the iterates θ_n and θ_{n+1}. It scales the current prediction error $(y_n - \varphi_n'\theta_n)$ to determine how much to adjust for the next estimate. We henceforth refer to μ as the *adaptation rate* of the algorithm. An important consideration for the adaptive filtering problem is the interplay between the transition rate ε of how fast the true system parameter α_n jumps and the adaptation rate μ of how quickly the estimates θ_n can adjust.

In [29], the LMS algorithm (2.2) was analyzed for the Markovian adaptive filtering problem 2.1 under the assumption that $\varepsilon = O(\mu)$; i.e. the adaptation rate of the estimates is nearly the same as the transition rate of the Markov chain.

In this chapter, we analyze the so-called Sign-Regressor algorithm for estimating the time-varying system parameter α_n which evolves as a Markov chain. In what follows, denote $\text{sgn}(y) = \mathcal{I}_{\{y>0\}} - \mathcal{I}_{\{y<0\}}$ for a scalar $y \in \mathbb{R}$, where $\mathcal{I}_{\{\cdot\}}$ is the indicator function of a set. For a vector $\phi = [\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(r)}] \in \mathbb{R}^r$, denote $\text{Sgn}(\phi) = [\text{sgn}(\phi^{(1)}), \ldots, \text{sgn}(\phi^{(r)})]$.

Algorithm 2. *The Sign-Regressor (SR) algorithm generates estimates θ_n recursively by the scheme*

$$\theta_{n+1} = \theta_n + \mu \mathrm{Sgn}(\varphi_n)(y_n - \varphi_n'\theta_n). \tag{2.3}$$

In many of the applications of adaptive filtering it is desirable to speed computations in order to effectively track the parameter. This is especially true in communication networks, when computations have to be carried out on-line with high dimensional data, frequent data shuffling, and limited resources. One method of speeding computations is to reduce the complexity of the data in the estimation scheme. In [11] a variant of the LMS algorithm (2.2), was proposed which uses only the sign of the residuals $(y_n - \varphi_n'\theta_n)$ to update the algorithm, i.e. $\theta_{n+1} = \theta_n + \mu\varphi_n \mathrm{sgn}(y_n - \varphi_n'\theta_n)$. This algorithm is now often referred to as the Sign-Error (SE) algorithm. Because of the $\mathrm{sgn}(\cdot)$ operator on the residuals, computations are reduced to simple bit shifts and the speed is substantially improved from the LMS algorithm. However, the highly non-linear operator $\mathrm{sgn}(\cdot)$ on the residuals makes analysis of the Sign-Error algorithm very difficult (this will be considered in Chapter 3). In addition, by 'throwing away' much of the information in the residuals, estimates from the SE algorithm tend to converge more slowly than the LMS algorithm.

The Sign-Regressor algorithm given in (2.3) can be thought of as a compromise between the LMS algorithm and the SE algorithm. By keeping the entirety of the information of the residuals $y_n - \varphi_n'\theta_n$ the SR modulates its adaptation by the magnitude of the current error (and not just the direction) and tends to converge at similar rates to the LMS algorithm. Instead the SR algorithm 'clips' the direction information from the regression vectors φ_n with the $\mathrm{Sgn}(\cdot)$ operator and still shows improved computation speed compared the the LMS algorithm (especially with large r for high-dimensional models). In addition, the linear form of the algorithm on the residuals makes the analysis much simpler from what we shall see in Chapter 3.

In what follows we shall analyze the properties of the Sign-Regressor algorithm for the Markovian adaptive filtering problem modeled by (2.1). We shall make use of the following assumptions for the analysis.

A 2.1. *The system parameter process α_n is a discrete-time homogeneous Markov chain with state space $\mathcal{M} = \{a_1, a_2, \ldots, a_{m_0}\}, a_i \in \mathbb{R}^r$. In addition, there exists a small $\varepsilon > 0$ such that the transition probability matrix of α_n is given by*

$$P^\varepsilon = I + \varepsilon Q \tag{2.4}$$

where I is the m_0-dimensional identity matrix and $Q = (q_{i,j}) \in \mathbb{R}^{m_0 \times m_0}$ is an irreducible generator of a continuous-time Markov chain, meaning that $q_{i,j} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m_0} q_{i,j} = 0$ for all i. Furthermore, assume the initial distribution $\pi_0 = [\mathbb{P}\{\alpha_0 = a_1\}, \mathbb{P}\{\alpha_0 = a_2\}, \ldots, \mathbb{P}\{\alpha_0 = a_{m_0}\}]$ is independent of ε.

A 2.2. *The sequences $\{\varphi_n\}, \{e_n\}$ are independent of the parameter process $\{\alpha_n\}$. Let $\mathcal{F}_n$ be the σ-algebra generated by $\{\varphi_j, e_j, \alpha_j : j < n; \alpha_n\}$ and let E_n be the conditional expectation with respect to $\mathcal{F}_n$. The sequence of signals $\{(\varphi_n, e_n)\}$ is bounded. In addition there exists a stable matrix $H \in \mathbb{R}^{r \times r}$ and a constant $K > 0$ such that for all n*

$$
\begin{aligned}
&\left| \sum_{j=n}^{\infty} E_n \left[\mathrm{Sgn}(\varphi_j)\varphi_j' - H \right] \right| \leq K \\
&\left| \sum_{j=n}^{\infty} E_n \left[\mathrm{Sgn}(\varphi_j)e_j \right] \right| \leq K
\end{aligned}
\tag{2.5}
$$

A 2.3. *For the matrix H as in A2.1 and for each $m \in \mathbb{N}$, as $n \to \infty$*

$$
\begin{aligned}
\frac{1}{n} \sum_{k=m}^{m+n} \mathrm{Sgn}\varphi_k \varphi_k' &\xrightarrow{p} H \\
\frac{1}{n} \sum_{k=m}^{m+n} \mathrm{Sgn}\varphi_k e_k &\xrightarrow{p} 0
\end{aligned}
\tag{2.6}
$$

Remark 2.1. We pause here to discuss the practicality of the assumptions. Assumption A2.1 formally defines the transition properties of the slow Markov chain α_n. We assume for simplicity that the underlying generator Q is a constant matrix, but time-varying $Q(t)$ can be treated as in [31]. We point out that for implementation of the algorithm neither Q nor the transition rate ε need be known.

Assumptions A2.2 and A2.3 on the signals $\{(\varphi_n, e_n)\}$ are quite broad. The conditions given in (2.5) and (2.6) characterize the signals as mixing processes and allow us to work with correlated signals whose distant future and distant past are asymptotically independent. The matrix H is then asymptotic covariance of the matrix $\mathrm{Sgn}(\varphi_n)\varphi_n'$. By H is stable we mean the its eigenvalues have negative real parts (H is a Hurwitz matrix).

The boundedness assumption of the signals is taken for simplicity of notation and can be removed in several ways. For example, one can use a truncation device on the estimates θ_n [19, Section 5.1] with randomly increasing bound as in [5]. Given a sequence of increasing truncation bounds, at any time instance, we compare the iterate computed with the truncation bound. If the iterate is larger than the bound, the truncation device forces the iterate to return to a bounded region, and the truncation bound is also updated; otherwise, the iterate is as without truncations. suppose. More explicitly, let $\{M_n\}$ be a monotone increasing sequence of positive real numbers such that $M_n \to \infty$ as $n \to \infty$. Define a sequence of nonnegative random variables $\{\varpi_n\}$ and the truncation algorithm as

$$
\begin{aligned}
\varpi_n &= \sum_{i=0}^{n-1} \mathcal{I}_{\left\{|\theta_i + \mu \mathrm{Sgn}(\varphi_i)(y_i - \varphi_i'\theta_i)| > M_{\varpi_i}\right\}}, \\
\theta_{n+1} &= [\theta_n + \mu \mathrm{Sgn}(\varphi_n)(y_n - \varphi_n'\theta_n)]\mathcal{I}_{\{|\theta_n + \mu \mathrm{Sgn}(\varphi_n)(y_n - \varphi_n'\theta_n)| \leq M_{\varpi_n}\}}.
\end{aligned}
\tag{2.7}
$$

Using the methods in [5], it can be shown that the above expanding random truncations are only executed a finite number of times, so eventually the algorithm will be bounded with probability one. With Markovian parameters, we can use the above truncations together with the methods to be used in this work to carry out the analysis. However, to ease the

already complex notation we shall assume regressors φ_n and errors e_n are bounded.

In practice, one chooses an appropriate adaptation rate μ without knowledge of the underlying transition rate ε. Depending on the relationship of ε to μ, one will see very different behavior in the limit system. We shall break down the limit analysis into three cases as follows.

(i) $\varepsilon = O(\mu)$ "On-Line" : The transition rate ε is on par with the adaptation rate μ, so the parameter α can jump about as quickly as θ can track it. The limit dynamics have occasional jumps in α_n, but the estimates θ_n are still able to track it closely.

(ii) $\varepsilon \ll \mu$ "Slower Markov Chain" : The transition rate for α_n is much slower than the adaptation rate for θ_n. More precisely, we shall assume $\varepsilon = O(\mu^{1+\eta})$ for some $0 < \eta \le 1$. In this case since the parameter α_n jumps so infrequently it is much as though α_n were constant, and the limit behavior is largely determined by the initial distribution.

(iii) $\varepsilon \gg \mu$ "Fast Markov Chain" : The transition rate for α_n is much faster than the adaptation rate for the estimates θ_n. More precisely, $\varepsilon = O(\mu^{\gamma})$ for some $1/2 \le \gamma < 1$. In this case the parameter α_n jumps too quickly for the estimate to track it. However, the frequent jumping of α means that it quickly comes to the stationary distribution $\nu = [\nu_1, \ldots, \nu_{m_0}]$ associated with the continuous-time generator Q, as does the distribution of the estimates.

2.2 Mean Squares Error Bounds

Let $\widetilde{\theta}_n \triangleq \theta_n - \alpha_n$ be the sequence of tracking errors for the estimates. We begin our analysis by establishing expected error bounds on $\mathbb{E}|\widetilde{\theta}_n|^2$ in terms of the adaptation rate μ of the algorithm and the transition rate ε of the Markov chain.

Theorem 2.2. *Under assumptions A2.1 and A2.2, there exists $N_{\mu,\varepsilon} > 0$ such that for all*

$n > N_{\mu,\varepsilon}$ *we have*

$$\mathbb{E}|\widetilde{\theta}_n|^2 = \mathbb{E}\,|\theta_n - \alpha_n|^2 = O\left(\mu + \varepsilon + \varepsilon^2/\mu\right). \tag{2.8}$$

Proof. Note that $\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \mu\mathrm{Sgn}(\varphi_n)(-\widetilde{\theta}_n + e_n)$. Define a Liapunov function $V(x) = (x'x)/2$. Then consider

$$\begin{aligned}
E_n V(\widetilde{\theta}_{n+1}) - V(\widetilde{\theta}_n) &= E_n \widetilde{\theta}'_n[-\mu\mathrm{Sgn}(\varphi_n)\varphi'_n\widetilde{\theta}_n + \mu\mathrm{Sgn}(\varphi_n)e_n + (\alpha_n - \alpha_{n+1})] \\
&\quad + E_n| - \mu\mathrm{Sgn}(\varphi_n)\varphi'_n\widetilde{\theta}_n + \mu\mathrm{Sgn}(\varphi_n)e_n + (\alpha_n - \alpha_{n+1})|^2.
\end{aligned} \tag{2.9}$$

We note that $\mathcal{I}_{\{\alpha_n=a_i\}}$ is $\mathcal{F}_n$ measurable. In addition, because the Markov chain $\{\alpha_n\}$ is independent of the signals $\{(\varphi_n, e_n)\}$ and has transition matrix of the form given in (2.4) we can write

$$E_n(\alpha_n - \alpha_{n+1}) = \sum_{i=1}^{m_0}\left[a_i - \sum_{j=1}^{m_0} a_j(\delta_{ij} + \varepsilon q_{ij})\right]\mathcal{I}_{\{\alpha_n=a_i\}} = O(\varepsilon). \tag{2.10}$$

Similar estimates also yield

$$E_n|\alpha_n - \alpha_{n+1}|^2 = O(\varepsilon). \tag{2.11}$$

Since $|\widetilde{\theta}_n| = |\widetilde{\theta}_n| \cdot 1 \leq (|\widetilde{\theta}_n|^2 + 1)/2$, we have

$$O(\varepsilon)|\widetilde{\theta}_n| \leq O(\varepsilon)(V(\widetilde{\theta}_n) + 1). \tag{2.12}$$

Using that the signals $\{(\varphi_n, e_n)\}$ are bounded, we obtain

$$\begin{aligned}
E_n| - \mu\mathrm{Sgn}(\varphi_n)&\varphi'_n\widetilde{\theta}_n + \mu\mathrm{Sgn}(\varphi_n)e_n + (\alpha_n - \alpha_{n+1})|^2 \\
&= E_n|\alpha_n - \alpha_{n+1}|^2 + O(\mu^2 + \mu\varepsilon)(V(\widetilde{\theta}_n) + 1).
\end{aligned} \tag{2.13}$$

Then by applying (2.11), (2.12), and (2.13) we have

$$
\begin{aligned}
E_n V(\widetilde{\theta}_{n+1}) - V(\widetilde{\theta}_n) &= E_n \widetilde{\theta}_n'[-\mu \mathrm{Sgn}(\varphi_n)\varphi_n'\widetilde{\theta}_n + \mu \mathrm{Sgn}(\varphi_n)e_n + (\alpha_n - \alpha_{n+1})] \\
&\quad + E_n|\alpha_n - \alpha_{n+1}|^2 + O(\mu^2 + \mu\varepsilon)(V(\widetilde{\theta}_n) + 1)
\end{aligned}
\tag{2.14}
$$

We now use a perturbed Liapunov function approach (see [19]) to derive estimates for the terms in (2.14). Define perturbations of the Liapunov function by

$$
V_1^\mu(\widetilde{\theta}, n) = -\mu \sum_{j=n}^\infty E_n \widetilde{\theta}'(\mathrm{Sgn}(\varphi_j)\varphi_j' - H)\widetilde{\theta}, \quad V_2^\mu(\widetilde{\theta}, n) = \mu \sum_{j=n}^\infty \widetilde{\theta}' E_n \mathrm{Sgn}(\varphi_j)e_j
$$

$$
V_3^\varepsilon(\widetilde{\theta}, n) = \sum_{j=n}^\infty \widetilde{\theta}' E_n(\alpha_j - \alpha_{j+1}), \quad V_4^\varepsilon(n) = \sum_{j=n}^\infty E_n(\alpha_n - \alpha_{n+1})'(\alpha_j - \alpha_{j+1}).
$$

$$\tag{2.15}$$

By A2.2 we can obtain $\mu\left|\sum_{j=n}^\infty [E_n\mathrm{Sgn}(\varphi_j)\varphi_j' - H]\right||\widetilde{\theta}|^2 \le O(\mu)(V(\widetilde{\theta}) + 1)$, so we have

$$
|V_1^\mu(\widetilde{\theta}, n)| \le O(\mu)(V(\widetilde{\theta}) + 1).
\tag{2.16}
$$

Using similar methods we also obtain

$$
|V_2^\mu(\widetilde{\theta}, n)| \le O(\mu)(V(\widetilde{\theta}) + 1).
\tag{2.17}
$$

We note that for small ε the transition matrix $P^\varepsilon = I + \varepsilon Q$ is irreducible with a stationary distribution ν_ε, so there exists a N_ε such that $|(I+\varepsilon Q)^n - \mathbb{1}\nu_\varepsilon| \le K\varepsilon$ for $n > N_\varepsilon p$. Telescoping with the above gives $\sum_{j=n}^\infty |(I + \varepsilon Q)^{j+1-n} - (I + \varepsilon Q)^{j-n}| = O(\varepsilon)$ and so

$$
|V_3^\varepsilon(\widetilde{\theta}, n)| \le O(\varepsilon)(V(\widetilde{\theta}) + 1), \quad |V_4^\varepsilon(n)| = O(\varepsilon).
\tag{2.18}
$$

Going back to $V_1^{\mu}(\widetilde{\theta}, n)$, we note

$$
\begin{aligned}
E_n V_1^{\mu}(\widetilde{\theta}_{n+1}, n+1) - V_1^{\mu}(\widetilde{\theta}_n, n) &= E_n V_1^{\mu}(\widetilde{\theta}_{n+1}, n+1) - E_n V_1^{\mu}(\widetilde{\theta}_n, n+1) \\
&\quad + E_n V_1^{\mu}(\widetilde{\theta}_n, n+1) - V_n^{\mu}(\widetilde{\theta}_n, n).
\end{aligned}
\tag{2.19}
$$

Applying A2.2 to the first difference with $\widetilde{\theta}$ fixed, we obtain

$$
E_n V_1^{\mu}(\widetilde{\theta}_n, n+1) - V_1^{\mu}(\widetilde{\theta}_n, n) = \mu E_n \widetilde{\theta}_n'(\mathrm{Sgn}(\varphi_n)\varphi_n' - H)\widetilde{\theta}_n.
\tag{2.20}
$$

For the second term with the index fixed we have

$$
\begin{aligned}
E_n V_1^{\mu}(\widetilde{\theta}_{n+1}, n+1) - E_n V_1^{\mu}(\widetilde{\theta}_n, n+1) &= -\mu \sum_{j=n+1}^{\infty} E_n (\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)'[E_{n+1}\mathrm{Sgn}(\varphi_j)\varphi_j' - H]\widetilde{\theta}_{n+1} \\
&\quad - \mu \sum_{j=n+1}^{\infty} E_n \widetilde{\theta}_n'[E_{n+1}\mathrm{Sgn}(\varphi_j)\varphi_j' - H](\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)
\end{aligned}
\tag{2.21}
$$

In the same manner as (2.11) we obtain

$$
\begin{aligned}
E_n|\widetilde{\theta}_{n+1} - \widetilde{\theta}_n| &\leq \mu E_n|\mathrm{Sgn}(\varphi_n)\varphi_n'||\widetilde{\theta}_n| + \mu E_n|\mathrm{Sgn}(\varphi_n)e_n| + O(\varepsilon) \\
&= O(\mu)(V(\widetilde{\theta}_n) + 1) + O(\varepsilon).
\end{aligned}
\tag{2.22}
$$

Moreover,

$$
\begin{aligned}
\left| \mu \sum_{j=n+1}^{\infty} E_n \widetilde{\theta}_n'[E_{n+1}\mathrm{Sgn}(\varphi_j)\varphi_j' - H](\widetilde{\theta}_{n+1} - \widetilde{\theta}_n) \right| &\leq O(\mu^2 + \mu\varepsilon)(V(\widetilde{\theta}_n) + 1), \\
\left| \mu \sum_{j=n+1}^{\infty} E_n (\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)'[E_{n+1}\mathrm{Sgn}(\varphi_j)\varphi_j' - H]\widetilde{\theta}_{n+1} \right| &\leq O(\mu^2 + \mu\varepsilon)(V(\widetilde{\theta}_n) + 1).
\end{aligned}
\tag{2.23}
$$

Putting the above together we arrive at

$$
E_n V_1^{\mu}(\widetilde{\theta}_{n+1}, n+1) - V_1^{\mu}(\widetilde{\theta}_n, n) = \mu E_n \widetilde{\theta}_n'(\mathrm{Sgn}(\varphi_n)\varphi_n' - H)\widetilde{\theta}_n + O(\mu^2 + \mu\varepsilon)
\tag{2.24}
$$

Using methods similar to the estimate for $E_n V_1^\mu(\widetilde{\theta}_{n+1}) - V_1^\mu(\widetilde{\theta}_n, n)$, we also obtain

$$E_n V_2^\mu(\widetilde{\theta}_{n+1}, n+1) - V_2^\mu(\widetilde{\theta}_n, n) = -\mu E_n \widetilde{\theta}_n' \mathrm{Sgn}(\varphi_n) e_n + O(\mu^2 + \mu\varepsilon)(V(\widetilde{\theta}_n) + 1),$$

$$E_n[V_3^\varepsilon(\widetilde{\theta}_{n+1}, n+1) - V_3^\varepsilon(\widetilde{\theta}_n, n)] = -E_n \widetilde{\theta}_n'(\alpha_n - \alpha_{n+1}) + O(\varepsilon^2 + \mu^2)(V(\widetilde{\theta}_n) + 1), \quad (2.25)$$

$$E_n[V_4^\varepsilon(n+1) - V_4^\varepsilon(n)] = -E_n|\alpha_n - \alpha_{n+1}|^2 + O(\varepsilon^2)$$

The above estimates lead us to define

$$W(\widetilde{\theta}, n) = V(\widetilde{\theta}) + V_1^\mu(\widetilde{\theta}, n) + V_2^\mu(\widetilde{\theta}, n) + V_3^\varepsilon(\widetilde{\theta}, n) + V_4^\varepsilon(n). \quad (2.26)$$

The condition in A2.2 that H is a stable matrix gives the existence of a $\lambda > 0$ such that $\widetilde{\theta}' H \widetilde{\theta} \geq \lambda V(\widetilde{\theta})$ and so $-\mu \widetilde{\theta}' H \widetilde{\theta} - \mu O(\widetilde{\theta}) \leq -\mu \lambda V(\widetilde{\theta})$. With the above, we apply (2.14) to (2.26) with estimates (2.24)–(2.25) and arrive at

$$\begin{aligned}
E_n W(\widetilde{\theta}_{n+1}, n+1) - W(\widetilde{\theta}_n, n) &\leq -\mu \widetilde{\theta}_n' H \widetilde{\theta}_n + O(\mu^2 + \varepsilon^2) \\
&\leq -\lambda\mu V(\widetilde{\theta}_n) + O(\mu^2 + \varepsilon^2)(V(\widetilde{\theta}_n) + 1) \quad (2.27) \\
&\leq -\lambda\mu W(\widetilde{\theta}_n, n) + O(\mu^2 + \varepsilon^2)(W(\widetilde{\theta}_n, n) + 1).
\end{aligned}$$

We note that replacing $V(\widetilde{\theta}, n)$ by $W(\widetilde{\theta}, n)$ simply results in another $O(\mu\varepsilon) = O(\mu^2 + \varepsilon^2)$ term by the estimates given in (2.16) –(2.18).

Take sufficiently small μ and ε such that there exists $0 < \lambda_0 \leq \lambda$ with $-\lambda\mu + O(\mu^2) + O(\varepsilon^2) \leq -\lambda_0\mu$. We then have the recursive inequality $E_n W(\widetilde{\theta}_{n+1}, n+1) \leq (1-\lambda_0\mu)W(\widetilde{\theta}_n, n) + O(\mu^2 + \varepsilon^2)$ and so by taking expectation we obtain

$$\mathbb{E}W(\widetilde{\theta}_{n+1}, n+1) \leq (1 - \lambda_0\mu)^{n-N_\varepsilon} \mathbb{E}W(\widetilde{\theta}_{N_\varepsilon}, N_\varepsilon) + O\left(\mu + \varepsilon^2/\mu\right) \quad (2.28)$$

One can now see that there exists $N_{\mu,\varepsilon}$ such that for $n \geq N_{\mu,\varepsilon}$ we have $(1 - \lambda_0\mu)^{n-N_\varepsilon} \leq O(\mu)$ and so $\mathbb{E}W(\widetilde{\theta}_{n+1}, n+1) \leq O(\mu + \varepsilon^2/\mu)$. To translate back to $V(\widetilde{\theta}_{n+1})$ we again apply (2.16)–(2.18) and finally obtain $EV(\widetilde{\theta}_{n+1}) \leq O(\mu + \varepsilon + \varepsilon^2/\mu)$ for $n \geq N_{\mu,\varepsilon}$. $\qquad \square$

2.3 Convergence Properties

We now consider the limit behavior of the sequence of estimates and true parameters $\{(\theta_n, \alpha_n)\}$. The analysis is carried out by examining a continuous-time interpolation of the discrete sequence. Define the μ-interpolated processes as

$$\theta^\mu(t) \stackrel{\Delta}{=} \theta_n, \quad \alpha^\mu(t) \stackrel{\Delta}{=} \alpha_n \qquad \text{for } t \in [n\mu, n\mu + \mu). \tag{2.29}$$

We examine the limit behavior at the infinitesimal level when $\mu \to 0$ by using weak convergence methods on the continuous-time interpolations $\theta^\mu(t)$ and $\alpha^\mu(t)$. Because we are interpolating the parameter process α at increments of μ while it in fact changes at rate ε, we shall see different limit behavior corresponding to the cases **(i)** $\varepsilon = O(\mu)$, **(ii)** $\varepsilon \ll \mu$, or **(iii)** $\varepsilon \gg \mu$.

For the cases $\varepsilon \ll \mu$ and $\varepsilon \gg \mu$, care must be taken since we are interpolating α_n by μ-increments while it changes at a rate of ε. This results in a two-time-scale Markov chain as in [31]. We will make use of the following calculation for the α limit behavior.

Remark 2.3. *Define a probability vector by $\pi_n^\varepsilon = (P(\alpha_n = a_1), \ldots, P(\alpha_n = a_{m_0})) \in \mathbb{R}^{1 \times m_0}$. Note that $\pi_0^\varepsilon = (\pi_{0,1}, \ldots, \pi_{0,m_0})$ (independent of ε). Because the Markov chain is time homogeneous, $(P^\varepsilon)^n$ is the n-step transition probability matrix with $P^\varepsilon = I + \varepsilon Q$. Then, for some $0 < \lambda_1 < 1$,*

$$\pi_n^\varepsilon = \pi(\varepsilon n) + O(\varepsilon + \lambda_1^{-n}), \quad 0 \le n \le O(1/\varepsilon), \tag{2.30}$$

where $\pi(t) = (\pi_1(t), \ldots, \pi_{m_0}(t))$ is the probability vector of the continuous Markov chain with generator Q which satisfies the Chapman-Kolmogorov equation

$$\frac{d}{dt}\pi(t) = \pi(t)Q, \quad \pi(0) = \pi_0. \tag{2.31}$$

We also obtain

$$(P^\varepsilon)^{n-n_0} = \Xi(\varepsilon n_0, \varepsilon n) + O(\varepsilon + \lambda_1^{-(n-n_0)}), \tag{2.32}$$

where with $t_0 = \varepsilon n_0$ and $t = \varepsilon n$, $\Xi(t_0, t)$ satisfies

$$\begin{cases} \dfrac{d}{dt}\Xi(t_0, t) = \Xi(t_0, t)Q, \\ \Xi(t_0, t_0) = I. \end{cases} \tag{2.33}$$

Consider the continuous-time interpolation of α_n by ε-increments (as opposed to μ in (2.29))

given as

$$\alpha^\varepsilon(t) := \alpha_n \quad for \ \ t \in [n\varepsilon, n\varepsilon + \varepsilon). \tag{2.34}$$

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

Q with state space $\mathcal{M}$. We can approximate $\mathbb{E}\alpha_n$ by

$$\mathbb{E}\alpha_n = \overline{\alpha}_*(\varepsilon n) + O(\varepsilon + \lambda_1^{-n}), \ \ for \ \ n \le O(1/\varepsilon),$$

$$\overline{\alpha}_*(\varepsilon n) \triangleq \sum_{j=1}^{m_0} a_j \pi_j(\varepsilon n). \tag{2.35}$$

The results obtained are in the sense of weak convergence. For a stochastic process X_n we shall write $X_n \xrightarrow{w} X$ to denote that X_n *converges weakly* to X, meaning that for any bounded and continuous function $f(\cdot)$, one has $\mathbb{E}f(X_n) \to f(X)$ as $n \to \infty$.

2.3.1 On-Line Limit $\varepsilon = O(\mu)$

We begin with the "On-Line" case $\varepsilon = O(\mu)$.

Theorem 2.4. *Let $\varepsilon = O(\mu)$, and assume A2.1, A2.2, and A2.3. Then as $\mu \to 0$, the processes $(\theta^\mu(t), \alpha^\mu(t))$ converges weakly to a process $(\theta(t), \alpha(t))$ (i.e. $(\theta^\mu, \alpha^\mu) \xrightarrow{w} (\theta, \alpha)$) such that $\alpha(t)$ is a continuous-time Markov chain generated by Q and $\theta(t)$ satisfies the Markov-switching ordinary differential equation*

$$\frac{d}{dt}\theta(t) = H(\alpha(t) - \theta(t)), \qquad \theta(0) = \theta_0 \tag{2.36}$$

We establish the theorem through a series of lemmas. For simplicity, we take $\varepsilon = \mu$

in what follows. We begin by using a truncation device to bound the estimates. Define $S_N \triangleq \{\theta \in \mathbb{R}^r : |\theta| \leq N\}$ to be the ball with radius N, and $q^N(\cdot)$ as a truncation function that is equal to 1 for $\theta \in S_N$, 0 for $\theta \in S_{N+1}$, and sufficiently smooth between. We then modify the algorithm 2.3 so that the estimates

$$\theta_{n+1}^N = \theta_n^N + \mu \text{Sgn}(\varphi_n)(y_n - \varphi_n'\theta_n)q^N(\theta_n^N) \tag{2.37}$$

are bounded by N. As before, interpolate by μ the discrete bounded estimates by $\theta^{N,\mu}(t) = \theta_n^N$ for $t \in [n\mu, n\mu + \mu)$.

To obtain the theorem, we shall first show that the sequence of bounded estimates and parameters $\{(\theta^{N,\mu}(\cdot), \alpha^\mu)\}_\mu$ is tight, thus allowing us to extract a weakly convergent subsequence by Prohorov's theorem. We then show that the limit sequence satisfies the Markov-switched differential equation. Lastly, we let the truncation bound $N \to \infty$ to show that the original sequence of estimates θ_n is also weakly convergent. For the following we shall write $D([0, \infty) : \mathbb{R}^r \times \mathcal{M})$ to denote the space of functions defined on $[0, \infty)$ taking values in $\mathbb{R}^r \times \mathcal{M}$ that are right continuous with left limits endowed with the Skorohod topology (see [19, Chapter 7] for definitions and further details).

Lemma 2.5. *The sequence $\{(\theta^{N,\mu}(\cdot), \alpha^\mu(\cdot))\}$ is tight in $D([0, \infty) : \mathbb{R}^r \times \mathcal{M})$.*

Proof. We have that $\{\alpha^\mu(\cdot)\}$ is tight by [31, Theorem 4.3] and that $\alpha^\mu(\cdot)$ converges weakly to a Markov chain generated by Q as noted in Remark 2.3. It remains to establish the the limit for the bounded estimate sequence $\{\theta^{N,\mu}(\cdot)\}_\mu$. We shall employ the tightness criterion given in [20, p.47], and so the goal is to show $\lim_{\delta \to 0} \limsup_{\mu \to 0} \left\{ \sup_{0 \leq s \leq \delta} \mathbb{E} \left| \theta^{N,\mu}(t + s) - \theta^{N,\mu}(t) \right|^2 \right\} = 0$, after which tightness shall be established.

We have that for any $\delta > 0$, and $t, s > 0$ satisfying $s \leq \delta$,

$$
\begin{aligned}
\mathbb{E} &\left| \theta^{N,\mu}(t+s) - \theta^{N,\mu}(t) \right|^2 \\
&\leq \mathbb{E} \left| \mu \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)(y_k - \varphi_k'\theta_k^N)q^N(\theta_k^N) \right|^2 \\
&\leq K\mathbb{E}\mu^2 \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)[\varphi_k'(\alpha_k - \theta_k^N)q^N(\theta_k^N) + e_k] \right|^2 \\
&\leq K\mathbb{E}\mu^2 \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)\varphi_k'(\alpha_k - \theta_k^N)q^N(\theta_k^N) \right|^2 + K\mathbb{E}\mu^2 \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)e_k \right|^2.
\end{aligned}
\tag{2.38}
$$

Applying the moment conditions on the signals $\operatorname{Sgn}(\varphi_k)\varphi_k'$ given in A2.2 with the boundedness of (θ_k^N, α_k) we have

$$
\begin{aligned}
E\mu^2 \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)\varphi_k'(\alpha_k - \theta_k^N)q^N(\theta_k^N) \right|^2 &\leq K\mu s \sum_{k=t/\mu}^{(t+s)/\mu-1} \mathbb{E}|\operatorname{Sgn}(\varphi_k)\varphi_k'|^2 \\
&\leq Ks^2 \leq K\delta^2.
\end{aligned}
\tag{2.39}
$$

and similarly

$$
\mathbb{E}\mu^2 \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} \operatorname{Sgn}(\varphi_k)e_k \right|^2 \leq K\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} \mathbb{E}|\operatorname{Sgn}(\varphi_k)e_k|^2 \leq Ks \leq K\delta.
\tag{2.40}
$$

Since the above estimates are uniform in μ we have

$$
\lim_{\delta \to 0} \limsup_{\mu \to 0} \left\{ \sup_{0 \leq s \leq \delta} \mathbb{E} \left| \theta^{N,\mu}(t+s) - \theta^{N,\mu}(t) \right|^2 \right\} = 0
$$

and hence $\{\theta^{N,\mu}(\cdot)\}$ is tight. $\quad\square$

With the tightness of $\{(\theta^{N,\mu}(\cdot), \alpha^\mu(\cdot))\}_\mu$ established, Prohorov's theorem implies that it is sequentially compact in the closure of $D([0,\infty) : \mathbb{R}^r \times \mathcal{M}$ equipped with the topology of weak convergence. Thus we shall extract such a weakly convergent subsequence and still denote it by $\{(\theta^{N,\mu}(\cdot), \alpha^\mu(\cdot))\}$ for notational simplicity. We write the limit as $(\theta^N(\cdot), \alpha(\cdot))$

and proceed to characterize the limit process. The following lemma is the main tool for the characterization and utilizes the weak convergence methods outlined in [19, Chapter 8].

Lemma 2.6. *Assume the A2.1 – A2.3. Then* $(\theta^{N,\mu}(\cdot), \alpha^{\mu}(\cdot)) \xrightarrow{w} (\theta^{N}(\cdot), \alpha(\cdot))$ *such that* $(\theta^{N}(\cdot), \alpha(\cdot))$ *is a solution of the martingale problem with operator*

$$\mathcal{L}_1^N f(\theta^N, i) = \nabla_\theta f(\theta^N, i) H[a_i - \theta^N] q^N(\theta^N) + \sum_{j=1}^{m_0} q_{ij} f(\theta^N, a_j), \tag{2.41}$$

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

$$f(\theta^N(t), \alpha(t)) - f(\theta^N(0), \alpha(0)) - \int_0^t \mathcal{L}_1^N f(\theta^N(\tau), \alpha(\tau)) d\tau$$

is a $\mathcal{F}_t$*-adapted martingale for each* $f \in C_0^1$ *and each* $a_j \in \mathcal{M}$.

Proof. As shown in [7, p.174], to derive the martingale limit we need only verify that for each C^1 function with compact support $f(\cdot, i)$, for each bounded and continuous function $h(\cdot)$, each $t, s > 0$, each positive integer κ, and each $t_i \leq t$ for $i \leq \kappa$,

$$\mathbb{E}h(\theta^N(t_i), \alpha(t_i) : i \leq \kappa) \Big[f(\theta^N(t+s), \alpha(t+s)) - f(\theta^N(t), \alpha(t))$$
$$- \int_t^{t+s} \mathcal{L}_1^N f(\theta^N(\tau), \alpha(\tau)) d\tau \Big] = 0. \tag{2.42}$$

For ease of notation we shall denote $h_N = h(\theta^N(t_i), \alpha(t_i) : i \leq \kappa)$ and $h_N^\mu = h(\theta^{N,\mu}(t_i), \alpha^\mu(t_i) : i \leq \kappa)$. Since $f(\cdot, i)$ is C_0^1 and $(\theta^{N,\mu}(\cdot), \alpha^\mu(\cdot)) \xrightarrow{w} (\theta^N(\cdot), (\alpha(\cdot))$ the Skorohod representation gives that as $\mu \to 0$,

$$\mathbb{E}h_N^\mu \left[f(\theta^{N,\mu}(t+s), \alpha^\mu(t+s)) - f(\theta^{N,\mu}(t), \alpha^\mu(t)) \right]$$
$$\to \mathbb{E}h_N \left[f(\theta^N(t+s), \alpha(t+s)) - f(\theta^N(t), \alpha(t)) \right]. \tag{2.43}$$

We subdivide the interval $[\frac{t}{\mu}, \frac{t+s}{\mu} - 1]$ by choosing an increasing sequence m_μ such that $m_\mu \to \infty$ as $\mu \to 0$ but $\delta_\mu \overset{\Delta}{=} \mu m_\mu \to 0$. The idea is that while the interval length $\frac{s}{\mu} \to \infty$ as $\mu \to 0$ we partition the interval into an increasing number m_μ of subintervals to approximate

the integral. However, we let the number of subdivisions m_μ grow slowly enough such that the subinterval length $\frac{s}{\mu m_\mu} = \frac{s}{\delta_\mu} \to \infty$ to allow for averaging. We then telescope over the endpoints of the subintervals and insert a term to examine changes in the estimate θ^N and the parameter α separately so that

$$
\begin{aligned}
\lim_{\mu \to 0} &\mathbb{E}h_N^\mu \left[f(\theta^{N,\mu}(t+s), \alpha^\mu(t+s)) - f(\theta^{N,\mu}(t), \alpha^\mu(t)) \right] \\
&= \lim_{\mu \to 0} \mathbb{E}h_N^\mu \left[\sum_{l\delta_\mu = t}^{t+s} [f(\theta_{lm_\mu + m_\mu}^N, \alpha_{lm_\mu + m_\mu}) - f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})] \right] \\
&= \lim_{\mu \to 0} \mathbb{E}h_N^\mu \Big[\sum_{l\delta_\mu = t}^{t+s} [f(\theta_{lm_\mu + m_\mu}^N, \alpha_{lm_\mu + m_\mu}) - f(\theta_{lm_\mu + m_\mu}^N, \alpha_{lm_\mu})] \\
&\qquad\qquad + \sum_{l\delta_\mu = t}^{t+s} [f(\theta_{lm_\mu + m_\mu}^N, \alpha_{lm_\mu}) - f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})] \Big].
\end{aligned}
\tag{2.44}
$$

We take the first order Taylor expansion of f at each endpoint of the subintervals lm_μ, and then telescope again through the iterates between the endpoints of the subinterval to obtain

$$
\begin{aligned}
\lim_{\mu \to 0} &\mathbb{E}h_N^\mu \Big[\sum_{l\delta_\mu = t}^{t+s} [f(\theta_{lm_\mu + m_\mu}^N, \alpha_{lm_\mu}) - f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})] \Big] \\
&= \lim_{\mu \to 0} \mathbb{E}h_N^\mu \sum_{l\delta_\mu = t}^{t+s} \delta_\mu \frac{1}{m_\mu} \Big[\sum_{k=lm_\mu}^{lm_\mu + m_\mu - 1} \nabla_\theta f(\theta_{lm_\mu}^N, \alpha_{lm_\mu}) \mathrm{Sgn}(\varphi_k) \varphi_k'(\alpha_k - \theta_k^N) q^N(\theta_k^N) \\
&\qquad\qquad + \sum_{k=lm_\mu}^{lm_\mu + m_\mu - 1} [\nabla_\theta f(\theta_{lm_\mu}^{N,+}, \alpha_{lm_\mu}) - \nabla_\theta f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})](\theta_{k+1}^N - \theta_k^N) q^N(\theta_k^N) \\
&\qquad\qquad\qquad\qquad + \sum_{k=lm_\mu}^{lm_\mu + m_\mu - 1} \nabla_\theta f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})] \mathrm{Sgn}(\varphi_k) e_k \Big],
\end{aligned}
\tag{2.45}
$$

where $\theta_{lm_\mu}^{N,+}$ is a point on the line segment joining $\theta_{lm_\mu}^N$ and $\theta_{lm_\mu + m_\mu}^N$. Since $\nabla_\theta f(\cdot, i)$ is continuous and $(\theta_{lm_\mu + \mu}^N - \theta_{lm_\mu}^N) \xrightarrow{w} 0$ as $\mu \to 0$ we have

$$
\lim_{\mu \to 0} \mathbb{E}h_N^\mu \sum_{k=lm_\mu}^{lm_\mu + m_\mu - 1} [\nabla_\theta f(\theta_{lm_\mu}^{N,+}, \alpha_{lm_\mu}) - \nabla_\theta f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})](\theta_{k+1}^N - \theta_k^N) q^N(\theta_k^N) = 0.
\tag{2.46}
$$

As $\mu \to 0$ we must have $\mu l m_\mu \to \tau \in [t, t+s]$. Thus for all k in the subinterval satisfying

$lm_\mu \le k \le lm_\mu + m_\mu - 1$, we have $\mu k \to \tau$ as well. Then considering the term involving the Markov chain α_k, since $\nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu})$ is $\mathcal{F}_{lm_\mu}$ measurable, we can insert the conditional expectation E_{lm_μ} to obtain

$$
\begin{aligned}
&\lim_{\mu \to 0} \mathbb{E} h^\mu_N \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \mathrm{Sgn}(\varphi_k)\varphi'_k \alpha_k \\
&= \lim_{\mu \to 0} \mathbb{E}^\mu_N \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) E_{lm_\mu}[\mathrm{Sgn}(\varphi_k)\varphi'_k - H]\alpha_k q^N(\theta^N_{lm_\mu}) \\
&\qquad\qquad + \lim_{\mu \to 0} \mathbb{E} h^\mu_N \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) E_{lm_\mu} H \alpha_k q^N(\theta^N_{lm_\mu}) \\
&= \lim_{\mu \to 0} \mathbb{E} h^\mu_N \sum_{l\delta_\mu=t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) H a_j \mathcal{I}_{\{\alpha_k=a_j\}} \\
&= \mathbb{E} h_N \int_t^{t+s} \nabla_\theta f(\theta^N(\tau), \alpha(\tau)) H \alpha(\tau) d\tau.
\end{aligned}
\tag{2.47}
$$

We note that we applied A2.3 to average out each sum $\frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} E_{lm_\mu}(\dots)$ since the number of iterates in each sub-sum $m_\mu \to \infty$ as $\mu \to 0$. Then taking the Riemann sum over intervals of length δ_μ, we obtained convergence to the integral. In a similar fashion we obtain

$$
\begin{aligned}
&\lim_{\mu \to 0} \mathbb{E} h^\mu_N \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \mathrm{Sgn}(\varphi_k)\varphi'_k \theta^N_k q^N(\theta^N_k) \\
&= \lim_{\mu \to 0} \mathbb{E} h^\mu_N \Big[\sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) E_{lm_\mu} H \theta^N_{lm_\mu} q^N(\theta^N_{lm_\mu}) \\
&\qquad + \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) E_{lm_\mu}[\mathrm{Sgn}(\varphi_k)\varphi'_k - H]\theta^N_{lm_\mu} q^N(\theta^N_{lm_\mu}) \Big] \\
&= \mathbb{E} h_N \int_t^{t+s} \nabla_\theta f(\theta^N(\tau), \alpha(\tau)) H \theta^N(\tau) d\tau,
\end{aligned}
\tag{2.48}
$$

as well as

$$
\lim_{\mu \to 0} \mathbb{E} h^\mu_N \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \mathrm{Sgn}(\varphi_k) e_k = 0. \tag{2.49}
$$

We use a similar process with the term of (2.45) which has $\theta^N_{(\cdot)}$ fixed and $\alpha_{(\cdot)}$ varying. By

exploiting the smooth and bounded property of f and applying Remark 2.3 we have that

$$
\begin{aligned}
\lim_{\mu \to 0} & \mathbb{E}h_N^\mu \sum_{l\delta_\mu=t}^{t+s} [f(\theta_{lm_\mu+m_\mu}^N, \alpha_{lm_\mu+m_\mu}) - f(\theta_{lm_\mu+m_\mu}^N, \alpha_{lm_\mu})] \\
& = \lim_{\mu \to 0} \mathbb{E}h_N^\mu \sum_{l\delta_\mu=t}^{t+s} [f(\theta_{lm_\mu}^N, \alpha_{lm_\mu+m_\mu}) - f(\theta_{lm_\mu}^N, \alpha_{lm_\mu})] + o(\mu) \\
& = \lim_{\mu \to 0} \mathbb{E}h_N^\mu \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \left[\frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} Qf(\theta_{lm_\mu}^N, \alpha_k)(\alpha_{lm_\mu}) \right] + o(\mu) \\
& = \mathbb{E}h_N \left[\int_t^{t+s} Qf(\theta^N(\tau), \alpha(\tau))d\tau \right].
\end{aligned}
$$

(2.50)

Thus by combining the above estimates (2.43)–(2.50), (2.42) is verified and the result follows.
□

Proof of Theorem 2.4. By Lemma 2.6, the solution $(\theta^N(\cdot), \alpha(\cdot))$ of the martingale problem with operator $\mathcal{L}_1^N$ satisfies the associated differential equation $\frac{d}{dt}\theta^N(t) = H[\alpha(t) - \theta^N(t)]q^N(\theta^N(t))$. It remains to show that as the truncation bound $N \to \infty$, the limit of the truncated sequence $\theta^N(\cdot)$ is the same as the limit of the untruncated sequence $\theta(\cdot)$.

Let $\mathcal{P}^0(\cdot)$ and $\mathcal{P}^N(\cdot)$ be the probability measures induced by $\theta(\cdot)$ and $\theta^N(\cdot)$ respectively. Since the solution of the differential equation associated with $\mathcal{L}_1^\mu$ is unique for each initial condition, the limit measure as $\mathcal{P}^0$ must be unique as well. For each finite time horizon $T < \infty$ the limit measure $\mathcal{P}^0(\cdot)$ must agree with $\mathcal{P}^N(\cdot)$ on all Borel paths in $D([0, \infty) : S_N)$. Then as $N \to \infty$, $\mathcal{P}^0\{\sup_{t \leq T} |\theta(t)| \leq N\} = 1$. Finally, the weak convergence $\theta^{N,\mu}(\cdot) \xrightarrow{w} \theta^N(\cdot)$ then gives $\theta^\mu(\cdot) \xrightarrow{w} \theta(\cdot)$.   □

We note that the Markov-switched limit

$$
\frac{d}{dt}\theta(t) = H[\alpha(t) - \theta(t)], \qquad \theta(t) = \theta_0
$$

is a novel feature of the analysis. While most results in classic stochastic approximation have a deterministic differential equation limit, the continuous-time Markov chain $\alpha(t)$ makes the

limit stochastic in nature. This Markov-switching ordinary differential equation limit is a special case of regime-switching diffusion models [32], which have recently gained popularity in many applications.

2.3.2 Slower Markov Chain: $\varepsilon \ll \mu$

We proceed now with the "Slower Markov Chain" case $\varepsilon \ll \mu$. In this case the Markov chain α_n transitions so slowly in relation to the adaptation rate that when we interpolate α by increments of μ the resulting process is essentially constant. Thus the limit dynamics are largely determined by the initial distribution $\pi_{0,i} = \mathbb{P}\{\alpha_0 = a_i\}$. In what follows write $\alpha_* = \sum_{i=1}^{m_0} a_i \pi_{0,i}$ for the mean of the Markov chain α against the initial distribution π_0. We then have the following result.

Theorem 2.7. *Let $\varepsilon = \mu^{1+\eta}$ for some $\eta > 0$, and assume A2.1, A2.2, and A2.3. Then as $\mu \to 0$, $\theta^\mu(\cdot) \xrightarrow{w} \theta(\cdot)$ such that $\theta(\cdot)$ is the solution of the ordinary differential equation*

$$\frac{d}{dt}\theta(t) = H\left(\alpha_* - \theta(t)\right), \qquad \theta(0) = \theta_0 \tag{2.51}$$

Proof. The proof is much the same as for Theorem 2.4. We shall only outline the key differences. Truncation is still used, but we omit the operator $q^N()$ for ease of notation. Tightness is obtained as before, and much of the estimates for the martingale limit remain the same. In the expansion of the term with the Markov chain we still see

$$\begin{aligned}
&\sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \mathrm{Sgn}(\varphi_k)\varphi_k'\alpha_k \\
&= \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [\mathrm{Sgn}(\varphi_k)\varphi_k' - H]\alpha_k + \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} H\alpha_k.
\end{aligned} \tag{2.52}$$

where the first term is averaged out in the limit. For the second term we note that from

Remark 2.3 for some $0 < \lambda_1 < 1$ we have,

$$
(P^\varepsilon)^{k-lm_\mu} = \Xi(\varepsilon k, \varepsilon lm_\mu) + O\left(\varepsilon + \lambda_1^{-(k-lm_\mu)}\right) \to I \quad \text{as} \quad \mu \to 0,
$$
$$
(P^\varepsilon)^{lm_\mu} = \Xi(0, \varepsilon lm_\mu) + O\left(\varepsilon + \lambda_1^{-lm_\mu}\right) \to I \quad \text{as} \quad \mu \to 0.
$$

Then we see

$$
\begin{aligned}
\mathbb{E}h_N^\mu &\sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta_{lm_\mu}, \alpha_{lm_\mu}) H\alpha_k \\
&= \mathbb{E}h_N^\mu \sum_{j_1=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \nabla_\theta f(\theta_{lm_\mu}, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} Ha_{j_1} E_{lm_\mu} \mathcal{I}_{\{\alpha_k=a_{j_1}\}} \\
&= \mathbb{E}h_N^\mu \sum_{j_1=1}^{m_0} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \nabla_\theta f(\theta_{lm_\mu}, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{i_1=1}^{m_0} \sum_{i_0=1}^{m_0} Ha_{j_1} \\
&\quad \times P(\alpha_k = a_{j_1}|\alpha_{lm_\mu} = a_{i_1}) P(\alpha_{lm_\mu} = a_{i_1}|\alpha_0 = a_{i_0}) P(\alpha_0 = a_{i_0}) \\
&\to \mathbb{E}h_N \sum_{i_0=1}^{m_0} \int_t^{t+s} \nabla_\theta f(\theta(\tau), \alpha(\tau)) Ha_{i_0} P(\alpha_0 = a_{i_0}) d\tau.
\end{aligned}
\tag{2.53}
$$

Other estimates are obtained similarly, and the rest of the proof follow as before with $H\alpha(t)$ replaced with $\sum_{i=1}^{m_0} Ha_i \pi_{0,i}$. $\qquad\square$

Given an initial distribution π_0, the limit against $\alpha_* = \sum_{i=1}^{m_0} a_i \pi_{0,i}$ is deterministic. In this case, we can obtain the following corollary.

Corollary 2.8. *Assume A2.1 – A2.3 and $\varepsilon = \mu^{1+\eta}$ for $1 < \eta \le 2$. Take any increasing sequence of time shifts $t_\mu \to \infty$ as $\mu \to 0$. Then $\theta^\mu(\cdot + t_\mu) \xrightarrow{w} \alpha_*$ as $\mu \to 0$.*

Proof. For any finite time horizon $T < \infty$ the pair $\{\theta^\mu(\cdot + t_\mu), \theta^\mu(\cdot + t_\mu - T)\}_\mu$ can be shown to be tight using the techniques in Theorem 2.2. Take a convergent subsequence with limit $(\theta(\cdot), \theta_T(\cdot))$ so that $\theta(0) = \theta_T(T)$. The value of $\theta_T(0)$ may not be known, but the set of possible $\{\theta_T(0)\}$ is tight since $\{\theta_n\}$ is tight. Consequently, we have

$$
\theta_T(T) = \exp(HT)\theta_T(0) - \int_0^T \exp(H(T-s))\alpha_* ds.
$$

Applying a change of variables $t = T - s$ in the right-hand side above, we arrive at

$$\theta_T(T) = \exp(GT)\theta_T(0) + \int_T^0 \exp(Ht)\alpha_* dt \to \alpha_* \quad \text{as} \ \ T \to \infty. \tag{2.54}$$

and the result thus follows. □

2.3.3 Fast Markov Chain: $\varepsilon \gg \mu$

Lastly, we consider the "Fast Markov Chain" case $\varepsilon \gg \mu$. Here the Markov Chain transitions much faster than adaptation rate μ. While the estimates are unable to track the parameter's all too frequent jumps, the large number of transitions allows the parameter process to quickly come to the stationary distbution ν associated with the underlying continuous time Markov chain. Write $\overline{\alpha} = \sum_{i=1}^{m_0} a_i \nu_i$ for the mean of the Markov chain α against the stationary distribution ν.

Theorem 2.9. *Let $\varepsilon = \mu^\gamma$ for some $1/2 < \gamma < 1$, and assume A2.1, A2.2, and A2.3. Then as $\mu \to 0$, $\theta^\mu(\cdot) \xrightarrow{w} \theta(\cdot)$ such that $\theta(\cdot)$ is the solution of the ordinary differential equation*

$$\frac{d}{dt}\theta(t) = H\left(\overline{\alpha} - \theta(t)\right), \qquad \theta(0) = \theta_0 \tag{2.55}$$

Proof. Again, since the technique is much the same as Theorem 2.4 we only present the key difference. When considering the integral limits $\mu l m_\mu \to \tau$ as $\mu \to 0$ we have that for $l m_\mu \le k < l m_\mu + m_\mu$ that $\varepsilon(k - l m_\mu) = \mu^\gamma(k - l m_\mu) \to \infty$. Thus applying Remark 2.3 we see

$$\Xi_{ij}(\varepsilon l m_\mu, \varepsilon k) = \nu_j + O(\varepsilon + \lambda_1^{-(k-l m_\mu)}),$$

and so

$$
\lim_{\mu \to 0} \mathbb{E} h_N^{\mu} \sum_{i_1=1}^{m_0} \sum_{\substack{l\delta_\mu=t}}^{t+s} \delta_\mu \nabla_\theta f(\theta_{lm_\mu}, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} H a_{i_1} E_{lm_\mu} \mathcal{I}_{\{\alpha_k=a_{i_1}\}}
$$
$$
= \mathbb{E} h_N \int_t^{t+s} \nabla_\theta f(\theta(\tau), \alpha(\tau)) H \Big(\sum_{i_1=1}^{m_0} a_{i_1} \nu_{i_1} \Big).
$$

(2.56)

The rest follows as before. $\qquad \square$

We again exploit the deterministic limit with $\bar{\alpha}$ to obtain the following corollary. The proof is the same as Corollary 2.8.

Corollary 2.10. *Assume A2.1 – A2.3 and $\varepsilon = \mu^\gamma$ for $0 < \gamma \le 1/2$. Then for any $t_\mu \to \infty$ as $\mu \to 0$, $\theta^\mu(\cdot + t_\mu) \xrightarrow{w} \bar{\alpha}$ as $\mu \to 0$.*

2.4 Asymptotic Distribution

Given that the process (θ_n, α) converges to a limit, one wishes to establish the rate at which the process converges. For adaptive algorithms with constant step-sizes μ the rate of convergence given by the appropriate scaling factor γ large enough such that the scaled error $(\theta_n - \alpha)/\mu^\gamma$ converges to a limit, yet small enough such that the limit is non-trivial.

Considering the result of Theorem 2.2, after interpolating at rate μ one expects the appropriate scaling factor to be $\gamma = 1/2$. We shall make use of the following assumption in this section.

A 2.4. *The scaled signals $\sqrt{\mu} \sum_{j=0}^{t/\mu-1} \mathrm{Sgn}(\varphi_j) e_j \xrightarrow{w} \widetilde{w}$, where $\widetilde{w}(t)$ is a Brownian motion with variance $\widetilde{\Sigma} t \widetilde{\Sigma} \in \mathbb{R}^{r \times r}$ positive definite.*

The above is a condition on the input and error signals $\{(\varphi_n, e_n)\}$ and is quite general. For example, suppose $\mathrm{Sgn}(\varphi_n) e_n = \varpi_n$ is a stationary mixing process with $\sum_n \widetilde{\phi}_n^{1/2} < \infty$, where $\widetilde{\phi}_n$ is the associated mixing measure. Then $\sqrt{\mu} \sum_{j=0}^{(t/\mu)-1} \varpi_j$ converges weakly to a Brownian motion $\widetilde{w}(t)$ with covariance $\widetilde{\Sigma} t$ such that the covariance $\widetilde{\Sigma}$ is given by $\widetilde{\Sigma} = \mathbb{E} \varpi_0 \varpi_0' + \sum_{j=1}^{\infty} \mathbb{E} \varpi_j \varpi_0' + \sum_{j=1}^{\infty} \mathbb{E} \varpi_0 \varpi_j'$. Further details can be found in [3].

We again begin with the case $\varepsilon = O(\mu)$. For simplicity we take $\varepsilon = \mu$ in what follows. By virtue of Theorem 2.2 there exists $N_{\mu,\varepsilon} = N_\mu$ such that $\mathbb{E}|\theta_n - \alpha_n|^2 = O(\mu)$. Then consider the scaled error

$$u_n = \frac{\theta_n - \alpha_n}{\sqrt{\mu}},$$
$$u^\mu(t) = u_n \quad \text{for } t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu) \tag{2.57}$$

so that $\mathbb{E}|u_n|^2 = O(1)$ for $n \geq N_\mu$, giving that $\{u_n : n \geq N_\mu\}$ is tight. Then we obtain the following theorem.

Theorem 2.11. *Let $\varepsilon = O(\mu)$ and assume A2.1 – A2.4. Then $u^\mu(\cdot) \xrightarrow{w} u(\cdot)$ such that $u(\cdot)$ is a solution to the stochastic differential equation*

$$du = Hu\,dt + \widetilde{\Sigma}^{1/2}dw \tag{2.58}$$

where $w(\cdot) \in \mathbb{R}^r$ is a standard Brownian motion.

Proof. The technique is similar as that in Section 2.3. Since the drift and diffusion coefficients are linear, there is a unique solution $u(t)$ to (2.58). The tightness of $\{u^\mu(\cdot)\}$ is argued above, and the weak limit is is shown to be u by establishing that the limit must solve the martingale problem with operator

$$\mathcal{L}f(u) \triangleq \nabla f(u)Hu + \frac{1}{2}\mathrm{tr}(\widetilde{\Sigma}\nabla^2 f(u)),$$

where $\nabla^2 f(u)$ is the Hessian of with respect to u. Estimates as in Theorem 2.4 with second order expansions are used to establish the result. See [19, Chapter 10] for detailed examples of the technique. Additionally, the analogous result under more difficult conditions is proven for the Sign-Error algorithm in Chapter 3. □

In the case $\varepsilon \ll \mu$, we consider the scaled error from α_*. We take $\varepsilon = \mu^{1+\eta}$ for $1 < \eta \leq 2$.

Define

$$v_n = \frac{\theta_n - \alpha_*}{\sqrt{\mu}} \qquad v^\mu(t) = v_n \quad \text{for } t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu) \qquad (2.59)$$

where $N_\mu = N_{\mu,\mu^{1+\eta}}$ is as in Theorem 2.2. We again omit the details in favor of brevity, but present the main result.

Theorem 2.12. *Let $\varepsilon = \mu^{1+\eta}$ for some $\eta > 0$ and assume A2.1 – A2.4. Then $v^\mu(\cdot) \xrightarrow{w} v(\cdot)$ such that $v(\cdot)$ is a solution to the stochastic differential equation*

$$dv = Hv\,dt + \widetilde{\Sigma}^{1/2}dw \qquad (2.60)$$

where $w(\cdot) \in \mathbb{R}^r$ is a standard Brownian motion.

Finally, for $\varepsilon \gg \mu$ we consider the scaled error from the stationary mean $\overline{\alpha}$. Take $\varepsilon = \mu^\gamma$ for some $1/2 \leq \gamma < 1$, and define

$$z_n = \frac{\theta_n - \alpha_*}{\sqrt{\mu}} \qquad z^\mu(t) = z_n \quad \text{for } t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu). \qquad (2.61)$$

Then the limit is as follows.

Theorem 2.13. *Let $\varepsilon = \mu^\gamma$ for $1/2 \leq \gamma < 1$ and assume A2.1 – A2.4. Then $z^\mu(\cdot) \xrightarrow{w} z(\cdot)$ such that $z(\cdot)$ is a solution to the stochastic differential equation*

$$dz = Hz\,dt + \widetilde{\Sigma}^{1/2}dw \qquad (2.62)$$

where $w(\cdot) \in \mathbb{R}^r$ is a standard Brownian motion.

Theorems 2.11, 2.12, and 2.13 characterize errors $\theta_n - \alpha_n$, $\theta_n - \alpha_*$, and $\theta_n - \overline{\alpha}$ respectively. For each case the theorems imply the asymptotic error is mean 0 with variance μS, where S is the solution to the Lyapunov equation $HS + SH' = -\widetilde{\Sigma}$.

3   Sign-Error Algorithms for Markovian Parameters

We now consider the Sign-Error (SE) algorithm for the adaptive filtering problem.

Algorithm 3. *The Sign-Error (SE) algorithm generates estimates θ_n recursively by the scheme*

$$\theta_{n+1} = \theta_n + \mu\varphi_n \mathrm{sgn}(y_n - \varphi_n'\theta_n). \tag{3.1}$$

Here the $\mathrm{sgn}(\cdot)$ operator is taken on the residuals $y_n - \varphi_n'\theta_n$. With an appropriate choice of the input "training" sequence, computations are reduced to bit shifts and the SE algorithm is able to be carried out with significant improvement in speed from the LMS and even SR algorithms. However, the hard operator on the residuals makes the analysis significantly more difficult than the LMS and SR algorithms. To obtain the desired results, we shall need slightly stronger conditions than was used for the SR algorithm.

A 3.1. *The system parameter process α_n is a discrete-time homogeneous Markov chain with state space $\mathcal{M} = \{a_1, a_2, \ldots, a_{m_0}\}, a_i \in \mathbb{R}^r$. In addition, there exists a small $\varepsilon > 0$ such that the transition probability matrix of α_n is given by*

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

where I is the m_0-dimensional identity matrix and $Q = (q_{i,j}) \in \mathbb{R}^{m_0 \times m_0}$ is an irreducible generator of a continuous-time Markov chain, meaning that $q_{i,j} \geq 0$ for $i \neq j$ and $\sum_{j=1}^{m_0} q_{i,j} = 0$ for all i. The initial distribution $\pi_0 = [\mathbb{P}\{\alpha_0 = a_1\}, \mathbb{P}\{\alpha_0 = a_2\}, \ldots, \mathbb{P}\{\alpha_0 = a_{m_0}\}]$ is independent of ε.

A 3.2. *The sequence of signals $\{(\varphi_n, e_n)\}$ is stationary and independent of the parameter process $\{\alpha_n\}$. The input signals $\{\varphi_n\}$ are taken to be uniformly bounded and $\{e_k\}$ is zero-mean. Let $\mathcal{F}_n$ be the σ-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.*

A 3.3. *For each $i = 1, \ldots, m_0$, define*

$$g_n \triangleq \varphi_n \mathrm{sgn}(\varphi'_n[\alpha_n - \theta_n] + e_n)$$

$$g_n(\theta, i) \triangleq \varphi_n \mathrm{sgn}(\varphi'_n[a_i - \theta] + e_n)\mathcal{I}_{\{\alpha_n = a_i\}} \qquad (3.3)$$

$$\widetilde{g}_n(\theta, i) \triangleq E_n g_n(\theta, i)$$

A 3.4. *For each n and i, there is an $A_n^{(i)} \in \mathbb{R}^{r \times r}$ such that given $\alpha_n = a_i$,*

$$\widetilde{g}_n(\theta, i) = A_n^{(i)}(a_i - \theta)\mathcal{I}_{\{\alpha_n = a_i\}} + o(|a_i - \theta|\mathcal{I}_{\{\alpha_n = a_i\}})$$

$$\mathbb{E}A_n^{(i)} = A^{(i)} \qquad (3.4)$$

Remark 3.1. Let us take a moment to justify the practicality of the assumptions. A3.1 is exactly the same as in Chapter 2, characterizing the transition of the Markovian parameters. A3.2 is similar to that in Chapter 2. However, the boundedness assumption is only on the input signals $\{\varphi_n\}$, leaving the error signals $\{e_n\}$ to be quite general (e.g., Gaussian, etc.). This is facilitated by the natural truncation on the error by the sign-operator ($|\mathrm{sgn}(\varphi'_n \alpha_n + e_n)| \leq 1$). As for the boundedness assumption on φ_n, it can be removed by using a truncation device as outlined in the beginning of Chapter 2. Moreover, one may also accommodate unbounded random inputs by assuming them to be a martingale difference sequence; for example, see the treatment in [26]. In fact, the analysis is easier because the signals are uncorrelated. In A3.3, we consider that while $g_n(\theta, i)$ is not smooth w.r.t. θ, its conditional expectation $\widetilde{g}_n(\theta, i)$ can be a smooth function of θ. The condition (3.4) indicates that $\widetilde{g}_n(\theta, i)$ is locally (near a_i) linearizable. For example, this is satisfied if the conditional joint density of (φ_n, e_n) with respect to $\{\varphi_j, e_j, j < n, \varphi_n\}$ is differentiable with bounded derivatives; see [33] for more discussion. Finally, A3.4 is essentially a mixing condition which indicates that the remote past and distant future are asymptotically independent. Hence we may work with correlated signals as long as the correlation decays sufficiently quickly between iterates.

3.1   Mean Squares Error Bounds

With the stronger assumptions, we can obtain the same error bound as in Chapter 2. Here we define $\widetilde{\theta}_n := \alpha_n - \theta_n$ (the negation of the error sequence of Chapter 2) to more easily facilitate the analysis on the residuals $y_n - \varphi_n'\theta_n$.

Theorem 3.2. *Assume* A3.1 – A3.4 . *Then there is an* $N_{\mu,\varepsilon} > 0$ *such that for all* $n \geq N_\mu$,

$$\mathbb{E}|\widetilde{\theta}_n|^2 = \mathbb{E}|\alpha_n - \theta_n|^2 = O\left(\mu + \varepsilon + \varepsilon^2/\mu\right). \tag{3.5}$$

Proof. As before, define $V(x) = (x'x)/2$. Observe that

$$\widetilde{\theta}_{n+1} = \alpha_{n+1} - \theta_{n+1} = \widetilde{\theta}_n - \mu\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n) + (\alpha_{n+1} - \alpha_n), \tag{3.6}$$

so

$$E_n V(\widetilde{\theta}_{n+1}) - V(\widetilde{\theta}_n) = E_n\widetilde{\theta}_n'[(\alpha_{n+1} - \alpha_n) - \mu\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n)]$$
$$+E_n|(\alpha_{n+1} - \alpha_n) - \mu\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n)|^2. \tag{3.7}$$

By A3.2, the Markov chain α_n is independent of (φ_n, e_n) and $\mathcal{I}_{\{\alpha_n=a_i\}}$ is $\mathcal{F}_n$-measurable. Since the transition matrix is of the form $P^\varepsilon = I + \varepsilon Q$, we obtain

$$E_n(\alpha_{n+1} - \alpha_n) = \sum_{i=1}^{m_0} E_n(\alpha_{n+1} - a_i\big|\alpha_n = a_i)\mathcal{I}_{\{\alpha_n=a_i\}}$$
$$= \sum_{i=1}^{m_0}\left[\sum_{j=1}^{m_0} a_j(\delta_{ij} + \varepsilon q_{ij}) - a_i\right]\mathcal{I}_{\{\alpha_n=a_i\}} \tag{3.8}$$
$$= O(\varepsilon).$$

Similarly,

$$E_n|\alpha_{n+1} - \alpha_n|^2$$
$$= \sum_{j=1}^{m_0}\sum_{i=1}^{m_0} |a_j - a_i|^2\mathcal{I}_{\{\alpha_n=a_i\}}P(\alpha_{n+1} = a_j|\alpha_n = a_i) \tag{3.9}$$
$$= \sum_{j=1}^{m_0}\sum_{i=1}^{m_0} |a_j - a_i|^2\mathcal{I}_{\{\alpha_n=a_i\}}(\delta_{ij} + \varepsilon q_{ij}) = O(\varepsilon).$$

Note that $|\widetilde{\theta}_n| = |\widetilde{\theta}_n| \cdot 1 \le (|\widetilde{\theta}_n|^2 + 1)/2$, so

$$O(\varepsilon)|\widetilde{\theta}_n| \le O(\varepsilon)(V(\widetilde{\theta}_n) + 1). \tag{3.10}$$

Since the signals $\{(\varphi_n, e_n)\}$ are bounded, we have

$$
\begin{aligned}
E_n&|(\alpha_{n+1} - \alpha_n) - \mu\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n)|^2 \\
&= E_n|\alpha_{n+1} - \alpha_n|^2 + O(\mu^2 + \mu\varepsilon)[V(\widetilde{\theta}_n) + 1]
\end{aligned} \tag{3.11}
$$

Applying (3.11) to (3.7), we arrive at

$$
\begin{aligned}
E_nV&(\widetilde{\theta}_{n+1}) - V(\widetilde{\theta}_n) \\
&= -\mu E_n\widetilde{\theta}_n'\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n) + E_n\widetilde{\theta}_n'(\alpha_{n+1} - \alpha_n) \\
&\quad + E_n|\alpha_{n+1} - \alpha_n|^2 + O(\mu^2 + \mu\varepsilon)[V(\widetilde{\theta}_n) + 1]
\end{aligned} \tag{3.12}
$$

Note also that by A3.3,

$$
\begin{aligned}
\mu E_n&\widetilde{\theta}_n'\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n) \\
&= \mu\sum_{i=1}^{m_0} E_n\widetilde{\theta}_n'\varphi_n\mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n)\mathcal{I}_{\{\alpha_n = a_i\}} \\
&= \mu\sum_{i=1}^{m_0} E_n\widetilde{\theta}_n'A_n^{(i)}\widetilde{\theta}_n\mathcal{I}_{\{\alpha_n = a_i\}} + \mu o(\widetilde{\theta}_n) \\
&= \mu\sum_{i=1}^{m_0} \widetilde{\theta}_n'[A_n^{(i)} - A^{(i)}]\widetilde{\theta}_n\mathcal{I}_{\{\alpha_n = a_i\}} \\
&\quad + \mu\sum_{i=1}^{m_0} \widetilde{\theta}_n'A^{(i)}\widetilde{\theta}_n\mathcal{I}_{\{\alpha_n = a_i\}} + \mu o(\widetilde{\theta}_n).
\end{aligned} \tag{3.13}
$$

To treat the first three terms in (3.12), we define the following perturbed Lypunov func-

tions by

$$V_1^\mu(\widetilde{\theta}, n) \triangleq \sum_{j=n}^{\infty} \sum_{i=1}^{m_0} -\mu E_n \widetilde{\theta}'[A_j^{(i)} - A^{(i)}]\widetilde{\theta} \mathcal{I}_{\{\alpha_j = a_i\}}$$

$$V_2^\mu(\widetilde{\theta}, n) \triangleq \sum_{j=n}^{\infty} \widetilde{\theta}' E_n(\alpha_{j+1} - \alpha_j) \tag{3.14}$$

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

By virtue of A3.4, we have

$$|V_1^\mu(\widetilde{\theta}, n)| \leq \mu \sum_{i=1}^{m_0} K|\widetilde{\theta}|^2 \sum_{j=n}^{\infty} \phi^{1/2}(j-n) \leq O(\mu)[V(\widetilde{\theta}) + 1] \tag{3.15}$$

As noted in Chapter 2, the irreducibility of Q implies that of $I + \varepsilon Q$ for sufficiently small $\varepsilon > 0$. Thus there is an N_ε such that for all $n \geq N_\varepsilon$, $|(I + \varepsilon Q)^k - \mathbb{1}\nu_\varepsilon| \leq \lambda_c^k$ for some $0 < \lambda_c < 1$, where ν_ε denotes the stationary distribution associated with the transition matrix $I + \varepsilon Q$. Then the difference of the $j + 1 - n$ and $j - n$ step transition matrices is given by

$$(I + \varepsilon Q)^{j+1-n} - (I + \varepsilon Q)^{j-n}$$
$$= [(I + \varepsilon Q) - I](I + \varepsilon Q)^{j-n}$$
$$= [(I + \varepsilon Q) - I][(I + \varepsilon Q)^{j-n} - \mathbb{1}\nu_\varepsilon] + [(I + \varepsilon Q) - I]\mathbb{1}\nu_\varepsilon$$
$$= (\varepsilon Q)[(I + \varepsilon Q)^{j-n} - \mathbb{1}\nu_\varepsilon].$$

The forgoing line above follows from the fact $Q\mathbb{1} = 0$, hence $[(I + \varepsilon Q) - I]\mathbb{1}\nu_\varepsilon = 0$. Thus

$$\sum_{j=n}^{\infty} |I + \varepsilon Q)^{j+1-n} - (I + \varepsilon Q)^{j-n}| \leq O(\varepsilon) \sum_{j=n}^{\infty} \lambda_c^{j-n} = O(\varepsilon). \tag{3.16}$$

The forgoing estimates lead to $\sum_{j=n}^{\infty} E_n(\alpha_{j+1} - \alpha_j) = O(\varepsilon)$ and as a result

$$|V_2^\mu(\widetilde{\theta}, n)| \leq O(\varepsilon)(V(\widetilde{\theta}) + 1). \tag{3.17}$$

and similarly

$$|V_3^\mu(n)| = O(\varepsilon), \tag{3.18}$$

so all the perturbations can be made small.

Now, we note that

$$
\begin{aligned}
&E_n V_1^\mu(\widetilde{\theta}_{n+1}, n+1) - V_1^\mu(\widetilde{\theta}_n, n) \\
&= E_n V_1^\mu(\widetilde{\theta}_{n+1}, n+1) - E_n V_1^\mu(\widetilde{\theta}_n, n+1) \\
&\quad + E_n V_1^\mu(\widetilde{\theta}_n, n+1) - V_n^\mu(\widetilde{\theta}_n, n).
\end{aligned}
\tag{3.19}
$$

where

$$
\begin{aligned}
&E_n V_1^\mu(\widetilde{\theta}_n, n+1) - V_1^\mu(\widetilde{\theta}_n, n) \\
&= \mu \sum_{i=1}^{m_0} \widetilde{\theta}_n'[A_n^{(i)} - A^{(i)}]\widetilde{\theta}_n \mathcal{I}_{\{\alpha_n = a_i\}},
\end{aligned}
\tag{3.20}
$$

and

$$
\begin{aligned}
&E_n V_1^\mu(\widetilde{\theta}_{n+1}, n+1) - E_n V_1^\mu(\widetilde{\theta}_n, n+1) \\
&= \mu \sum_{j=n+1}^{\infty} \sum_{i=1}^{m_0} E_n(\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)'[A_j^{(i)} - A^{(i)}]\widetilde{\theta}_{n+1} \mathcal{I}_{\{\alpha_n = a_i\}} \\
&\quad + \mu \sum_{j=n+1}^{\infty} \sum_{i=1}^{m_0} E_n \widetilde{\theta}_n'[A_j^{(i)} - A^{(i)}](\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)\mathcal{I}_{\{\alpha_n = a_i\}}.
\end{aligned}
\tag{3.21}
$$

Using (3.8), we have

$$
\begin{aligned}
E_n|\widetilde{\theta}_{n+1} - \widetilde{\theta}_n| &\le E_n|\alpha_{n+1} - \alpha_n| + \mu E_n|\varphi_n \mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n)| \\
&\le O(\varepsilon + \mu).
\end{aligned}
\tag{3.22}
$$

Thus, in view of A3.4

$$
\begin{aligned}
&\left| \mu \sum_{j=n+1}^{\infty} \sum_{i=1}^{m_0} E_n \widetilde{\theta}_n' E_{n+1}[A_j^{(i)} - A^{(i)}](\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)\mathcal{I}_{\{\alpha_n = a_i\}} \right| \\
&\le O(\mu^2 + \mu\varepsilon)[V(\widetilde{\theta}_n) + 1],
\end{aligned}
\tag{3.23}
$$

and

$$
\begin{aligned}
&\left| \mu \sum_{j=n+1}^{\infty} \sum_{i=1}^{m_0} E_n(\widetilde{\theta}_{n+1} - \widetilde{\theta}_n)' E_{n+1}[A_j^{(i)} - A^{(i)}]\widetilde{\theta}_{n+1} \mathcal{I}_{\{\alpha_n = a_i\}} \right| \\
&\le O(\mu^2 + \mu\varepsilon)[V(\widetilde{\theta}_n) + 1].
\end{aligned}
\tag{3.24}
$$

Putting together (3.19)–(3.24), we establish that

$$
\begin{aligned}
E_n V_1^\mu(\widetilde\theta_{n+1}, n+1) - V_1^\mu(\widetilde\theta_n, n) = \\
\mu \sum_{i=1}^{m_0} E_n \widetilde\theta_n'[A_j^{(i)} - A^{(i)}]\widetilde\theta_n \mathcal{I}_{\{\alpha_n=a_i\}} + O(\mu^2 + \mu\varepsilon)[V(\widetilde\theta_n) + 1].
\end{aligned}
\tag{3.25}
$$

Likewise, we can obtain

$$
\begin{aligned}
E_n V_2^\mu(\widetilde\theta_{n+1}, n+1) - V_2^\mu(\widetilde\theta_n, n) \\
= -E_n \widetilde\theta_n'(\alpha_{n+1} - \alpha_n) + O(\varepsilon^2 + \mu^2)
\end{aligned}
\tag{3.26}
$$

and

$$
E_n V_3^\mu(n+1) - V_3^\mu(n) = -E_n|\alpha_{n+1} - \alpha_n|^2 + O(\varepsilon^2).
\tag{3.27}
$$

Now we define

$$
W(\widetilde\theta, n) = V(\widetilde\theta) + V_1^\mu(\widetilde\theta, n) + V_2^\mu(\widetilde\theta, n) + V_3^\mu(n).
$$

Since each $A^{(i)}$ is a stable matrix there is a $\lambda > 0$ such that $\widetilde\theta' A^{(i)}\widetilde\theta \geq \lambda V(\widetilde\theta)$ for each i. Thus we may take λ such that $-\mu \sum_{i=1}^{m_0} \widetilde\theta' A^{(i)}\widetilde\theta \mathcal{I}_{\{\alpha_n=a_i\}} - \mu O(\widetilde\theta) \leq -\lambda\mu V(\widetilde\theta)$. Using this along with (3.7), (3.13), (3.25)–(3.27), and the inequality $O(\mu\varepsilon) = O(\mu^2 + \varepsilon^2)$, we arrive at

$$
\begin{aligned}
E_n W(\widetilde\theta_{n+1}, n+1) - W(\widetilde\theta_n, n) \\
= -\mu \sum_{i=1}^{m_0} \widetilde\theta_n' A^{(i)}\widetilde\theta_n \mathcal{I}_{\{\alpha_n=a_i\}} - \mu O(\widetilde\theta_n) + O(\mu^2 + \varepsilon^2)[V(\widetilde\theta_n) + 1] \\
\leq -\lambda\mu V(\widetilde\theta_n) + O(\mu^2 + \varepsilon^2)[V(\widetilde\theta_n) + 1] \\
\leq -\lambda\mu W(\widetilde\theta_n, n) + O(\mu^2 + \varepsilon^2)[W(\widetilde\theta_n, n) + 1].
\end{aligned}
\tag{3.28}
$$

Choose μ and ε small enough so that there is a $\lambda_0 > 0$ satisfying $\lambda_0 \leq \lambda$ and

$$
-\lambda\mu + O(\mu^2) + O(\varepsilon^2) \leq -\lambda_0\mu.
$$

Then we obtain

$$E_n W(\widetilde{\theta}_{n+1}, n+1) \le (1 - \lambda_0 \mu) W(\widetilde{\theta}_n, n) + O(\mu^2 + \varepsilon^2).$$

Note that there is an $N_\mu > 0$ such that $(1 - \lambda_0 \mu)^n \le O(\mu)$ for $n \ge N_\mu$. Taking expectation in the iteration for $W(\widetilde{\theta}_n, n)$ and iterating on the resulting inequality yield

$$\mathbb{E}W(\widetilde{\theta}_{n+1}, n+1) \le (1 - \lambda_0 \mu)^n W(\widetilde{\theta}_0, 0) + O\left(\mu + \varepsilon^2/\mu\right).$$

Thus

$$\mathbb{E}W(\widetilde{\theta}_{n+1}, n+1) \le O(\mu + \varepsilon^2/\mu).$$

Finally, applying (3.15)–(3.18) again, we also obtain

$$\mathbb{E}V(\widetilde{\theta}_{n+1}) \le O(\mu + \varepsilon + \varepsilon^2/\mu).$$

Thus the bound is established for μ and ε sufficiently small and $n \ge \max\{N_\mu, N_\varepsilon\} = N_{\mu,\varepsilon}$.
$\square$

3.2   Convergence Properties

As in Chapter 2, we investigate the limit of the estimate-parameter pair (θ_n, α_n) in three cases: $\varepsilon = O(\mu)$, $\varepsilon \ll \mu$, and $\varepsilon \gg \mu$. We again interpolate the discrete processes to continuous time as follows for the analysis.

$$\theta^\mu(t) \stackrel{\Delta}{=} \theta_n, \quad \alpha^\mu(t) \stackrel{\Delta}{=} \alpha_n \qquad \text{for } t \in [n\mu, n\mu + \mu). \tag{3.29}$$

On-Line: $\varepsilon = O(\mu)$

Beginning with the case $\varepsilon = O(\mu)$, we have the following result.

Theorem 3.3. *Let $\varepsilon = O(\mu)$ and assume A3.1 – A3.4. Then $(\theta^\mu(\cdot), \alpha^\mu(\cdot)) \xrightarrow{w} (\theta(\cdot), \alpha(\cdot))$ such that $\alpha(\cdot)$ is a continuous-time Markov chain with generator Q and $\theta(\cdot)$ satisfies the Markov-switched ODE*

$$\frac{d}{dt}\theta(t) = A^{(\alpha(t))}\left(\alpha(t) - \theta(t)\right), \qquad \theta(0) = \theta_0 \tag{3.30}$$

To obtain the limit we use the same techniques presented in Chapter 2. We begin by employing the truncation device $q^N(\cdot)$ bound the estimates. Define

$$\theta_{n+1}^N := \theta_n^N + \mu\varphi_n\mathrm{sgn}(y_n - \varphi_n'\theta_n^N)q^N(\theta_n^N), \ n = 0, 1, \dots, \\ \theta^{N,\mu}(t) := \theta_n^N \ \text{for} \ t \in [\mu n, \mu n + \mu). \tag{3.31}$$

As before, we begin by showing the limit is tight so as to extract a weakly convergent subsequence.

Lemma 3.4. *The sequence $(\theta^{N,\mu}(\cdot), \alpha^\mu(\cdot))$ is tight in $D([0,\infty) : \mathbb{R}^r \times \mathcal{M})$.*

Proof. By [31, Theorem 4.3], $\alpha^\mu(\cdot)$ is tight. As for $\theta^{N,\mu}(\cdot)$, we use the criterion given in [20, p.47]. With slight abuse of notation, denote $\mathcal{F}_t^\mu$ the σ-algebra generated by $\{(\varphi_j, e_j) : j \leq t/\mu\}$ and E_t^μ the respective conditional expectation. Then for any $\delta > 0$, and $t, s > 0$ satisfying $s \leq \delta$,

$$\begin{aligned}
&E_t^\mu\left|\theta^{N,\mu}(t+s) - \theta^{N,\mu}(t)\right|^2 \\
&\leq E_t^\mu\left|\mu\sum_{k=t/\mu}^{(t+s)/\mu-1}\varphi_k\mathrm{sgn}(y_k - \varphi_k'\theta_k^N)q^N(\theta_k^N)\right|^2 \\
&\leq \mu^2 E_t^\mu\sum_{j=t/\mu}^{(t+s)/\mu-1}\sum_{k=t/\mu}^{(t+s)/\mu-1}\varphi_j'\varphi_k\mathrm{sgn}(y_j - \varphi_j'\theta_j^N)q^N(\theta_j^N) \\
&\hspace{5cm}\times\mathrm{sgn}(y_k - \varphi_k'\theta_k^N)q^N(\theta_k^N) \\
&\leq \mu^2\sum_{j=t/\mu}^{(t+s)/\mu-1}\sum_{k=t/\mu}^{(t+s)/\mu-1}E_k\left|\varphi_j\right|^2 E_k\left|\varphi_k\right|^2 \\
&\leq O(s^2) \leq O(\delta^2)
\end{aligned} \tag{3.32}$$

uniformly in μ. Then

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

which establishes the criterion. $\quad\square$

Since $(\theta^{N,\mu}(\cdot),\alpha^\mu(\cdot))$ is tight, by Prohorov's theorem, we can extract a weakly convergence subsequence. Select such a subsequence and still denote it by $(\theta^{N,\mu}(\cdot),\alpha^\mu(\cdot))$ for notational simplicity, and write $(\theta^N(\cdot),\alpha(\cdot))$ for the limit. We characterize the limit with the following lemma.

Lemma 3.5. *The sequence* $(\theta^{N,\mu}(\cdot),\alpha^\mu(\cdot))\xrightarrow{w}(\theta^N(\cdot),\alpha(\cdot))$ *which solves the martingale problem with operator*

$$\mathcal{L}_1^N f(\theta^N,a_i):=\nabla f'(\theta^N,a_i)A^{(i)}[a_i-\theta^N]q^N(\theta^N)+\sum_{j=1}^{m_0}q_{ij}f(\theta^N,a_j),\qquad(3.33)$$

where for each $i\in\mathcal{M}$, $f(\cdot,i)\in C_0^1$.

Proof. While the technique is similar to that in Chapter 2, the details become more complicated because of the non-linear operation on the residuals in the SE algorithm. As before, to derive the martingale limit we show that for the C^1 function with compact support $f(\cdot,i)$, for each bounded and continuous function $h(\cdot)$, each $t,s>0$, each positive integer κ, and each $t_i\le t$ for $i\le\kappa$,

$$\mathbb{E}h(\theta^N(t_i),\alpha(t_i):i\le\kappa)\Big[f(\theta^N(t+s),\alpha(t+s))-f(\theta^N(t),\alpha(t))-\int_t^{t+s}\mathcal{L}_1^N f(\theta^N(\tau),\alpha(\tau))d\tau\Big]=0.$$
$$(3.34)$$

We shall use the notation $h^{N,\mu}\triangleq h(\theta^{N,\mu}(t_i),\alpha^\mu(t_i):i\le\kappa)$ and $h^N\triangleq h(\theta^N(t_i),\alpha(t_i):i\le\kappa)$. Since $f(\cdot,i)$ is smooth,

$$\lim_{\mu\to 0}\mathbb{E}h^{N,\mu}\Big[f(\theta^{N,\mu}(t+s),\alpha^\mu(t+s))-f(\theta^{N,\mu}(t),\alpha^\mu(t))\Big]$$
$$=\mathbb{E}h^N\Big[f(\theta^N(t+s),\alpha(t+s))-f(\theta^N(t),\alpha(t))\Big].$$
$$(3.35)$$

We use

$$\theta^{N,\mu}(t+s) - \theta^{N,\mu}(t) = \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu\varphi_k \text{sgn}(\varphi'_k[\alpha_k - \theta_k] + e_k)q^N(\theta_k^N) \tag{3.36}$$

to see that

$$
\begin{aligned}
&\lim_{\mu\to 0} \mathbb{E}h^{N,\mu}[f(\theta^{N,\mu}(t+s), \alpha^\mu(t+s)) - f(\theta^{N,\mu}(t), \alpha^\mu(t))]\\
&= \lim_{\mu\to 0} \mathbb{E}h^{N,\mu}\Big[\sum_{l\delta_\mu=t}^{t+s} [f(\theta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu}) - f(\theta^N_{lm_\mu}, \alpha_{lm_\mu})]\Big]\\
&= \lim_{\mu\to 0} \mathbb{E}h^{N,\mu}\Big[\sum_{l\delta_\mu=t}^{t+s} [f(\theta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu+m_\mu}) - f(\theta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu})]\\
&\qquad\qquad + \sum_{l\delta_\mu=t}^{t+s} [f(\theta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu}) - f(\theta^N_{lm_\mu}, \alpha_{lm_\mu})]\Big].
\end{aligned}
\tag{3.37}
$$

Working with the last term in (3.37) we use the Taylor expansion to obtain

$$
\begin{aligned}
&\lim_{\mu\to 0} \mathbb{E}h^{N,\mu}\Big[\sum_{l\delta_\mu=t}^{t+s} [f(\theta^N_{lm_\mu+m_\mu}, \alpha_{lm_\mu}) - f(\theta^N_{lm_\mu}, \alpha_{lm_\mu})]\Big]\\
&= \lim_{\mu\to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu=t}^{t+s} \Big[\delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu})\varphi_k \text{sgn}(\varphi'_k(\alpha_k - \theta_k^N) + e_k)q^N(\theta_k^N)\\
&\qquad + \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [\nabla_\theta f(\theta^{N,+}_{lm_\mu}, \alpha_{lm\mu}) - \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm\mu})](\theta_{k+1}^N - \theta_k^N)q^N(\theta_k^N)\Big].
\end{aligned}
\tag{3.38}
$$

where $\theta^{N,+}_{lm_\mu}$ is a point on the line segment joining $\theta^N_{lm_\mu}$ and $\theta^N_{lm_\mu+m_\mu}$. Since $|\theta^N_{lm_\mu+m_\mu} - \theta^N_{lm_\mu}| = O(\delta_\mu)$ and $\nabla_\theta f(\cdot, i)$ is smooth, we have the last term in (3.38) is $o(1)$ in the sense of probability as $\mu \to 0$. To work with the first term we insert the conditional expectation E_k and apply (3.4) to obtain

$$
\begin{aligned}
&\lim_{\mu\to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) E_k[\varphi'_k \text{sgn}(\varphi'_k(\alpha_k - \theta_k^N) + e_k)]q^N(\theta_k^N)\\
&= \lim_{\mu\to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu=t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \nabla f'(\theta^N_{lm_\mu}, \alpha_{lm_\mu})\\
&\qquad\qquad\qquad \times \Big[A_k^{(j)}(a_j - \theta_k^N) + o(|\alpha_k - \theta_k^N|)\Big] q^N(\theta_k^N) \mathcal{I}_{\{\alpha_k=a_j\}}.
\end{aligned}
\tag{3.39}
$$

Then

$$\mathbb{E}\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}\nabla f'(\theta_{lm_\mu}^N,\alpha_{lm_\mu})o(|\alpha_k-\theta_k^N|)\le K\mathbb{E}|\alpha_{lm_\mu}-\theta_{lm_\mu}|=O(\mu^{1/2}). \qquad (3.40)$$

Letting $\mu lm_\mu\to\tau$, then by (3.4),

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h^{N,\mu}\sum_{l\delta_\mu=t}^{t+s}\sum_{j=1}^{m_0}\frac{\delta_\mu}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}\Big[\nabla f'(\theta_{lm_\mu}^N,\alpha_{lm_\mu})A_k^{(j)}(a_j-\theta_k^N)q^N(\theta_k^N)\mathcal{I}_{\{\alpha_k=a_j\}}\Big]\\
&=\lim_{\mu\to0}\mathbb{E}h^{N,\mu}\sum_{l\delta_\mu=t}^{t+s}\sum_{j=1}^{m_0}\frac{\delta_\mu}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}\nabla f'(\theta_{lm_\mu}^N,\alpha_{lm_\mu})\Big[A^{(j)}(a_j-\theta_k^N)\\
&\hspace{6cm}+[A_k^{(j)}-A^{(j)}](a_j-\theta_k^N)\Big]q^N(\theta_k^N)\mathcal{I}_{\{\alpha_k=a_j\}}\\
&=\lim_{\mu\to0}\mathbb{E}h^{N,\mu}\sum_{l\delta_\mu=t}^{t+s}\sum_{j=1}^{m_0}\delta_\mu\nabla f'(\theta_{lm_\mu}^N,\alpha_{lm_\mu})\frac{1}{m_\mu}\sum_{k=lm_\mu}^{lm_\mu+m_\mu-1}A^{(j)}(a_j-\theta_k^N)q^N(\theta_k^N)\mathcal{I}_{\{\alpha_k=a_j\}}\\
&=\mathbb{E}h^N\int_t^{t+s}\nabla f'(\theta^N(\tau),\alpha(\tau))A^{(\alpha(\tau))}[\alpha(\tau)-\theta^N(\tau)]q^N(\tau)d\tau.
\end{aligned}$$
$$(3.41)$$

In a similar fashion, we obtain

$$\begin{aligned}
&\lim_{\mu\to0}\mathbb{E}h^{N,\mu}\sum_{l\delta_\mu=t}^{t+s}[f(\theta_{lm_\mu+m_\mu}^N,\alpha_{lm_\mu+m_\mu})-f(\theta_{lm_\mu+m_\mu}^N,\alpha_{lm_\mu})]\\
&=\mathbb{E}h^N\Big[\int_t^{t+s}Qf(\theta^N(\tau),\alpha(\tau))d\tau\Big].
\end{aligned}$$
$$(3.42)$$

Combining the above we verify ((3.34)), establishing the result of the lemma. □

Proof of Theorem 3.3. With Lemma 3.5, we have the truncated sequence $\theta^N(\cdot)$ satisfies the switched ODE $\dot\theta^N(t)=A^{(\alpha(t))}[\alpha(t)-\theta^N(t)q^N(t)]$, $\theta(0)=\theta_0$. Letting $N\to\infty$, showing that the limit of the untruncated sequence $\theta(\cdot)$ is the same as the limit of $\theta^N(\cdot)$ as $N\to\infty$ follows the same as the analogue in Chapter 2. The Theorem 3.3 follows. □

Slower Markov Chain: $\varepsilon\ll\mu$

In the case $\varepsilon\ll\mu$, the limit is again characterized by the initial distribution π_0 of α_0.

Theorem 3.6. *Let $\varepsilon = \mu^{1+\eta}$ for $0 < \eta \leq 1$ and assume A3.1 – A3.4. Then $\theta^\mu(\cdot) \xrightarrow{w} \theta(\cdot)$ such that $\theta(\cdot)$ is the solution to the ODE*

$$\frac{d}{dt}\theta(t) = \sum_{i=1}^{m_0} A^{(i)} (a_i - \theta(t)) \pi_{0,i}, \qquad \theta(0) = \theta_0 \tag{3.43}$$

Proof. We only note the key difference in the proof from the On-Line case $\varepsilon = \mu$ which results in the term involving the Markov chain α_n in (3.41). We see that

$$\lim_{\mu \to 0} Eh(\theta^{N,\mu}(t_i), \alpha^\mu(t_i) : i \leq \kappa) \sum_{l\delta_\mu = t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \frac{1}{m_\mu} A^{(j)} a_j I_{\{\alpha_k = a_j\}}$$

$$= \lim_{\mu \to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu = t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} A^{(j)} a_j E_{lm_\mu} I_{\{\alpha_k = a_j\}}$$

$$= \lim_{\mu \to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu = t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \nabla_\theta f(\theta^N_{lm_\mu}, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} \sum_{i_1=1}^{m_0} \sum_{i_0=1}^{m_0} A^{(j)} a_j P(\alpha_k = a_j | \alpha_{lm_\mu} = a_{i_1})$$

$$\times P(\alpha_{lm_\mu} = a_{i_1} | \alpha_0 = a_{i_0}) P(\alpha_0 = a_{i_0})$$

$$= \mathbb{E}h^N \sum_{i_0=1}^{m_0} \int_t^{t+s} \nabla_\theta f(\theta(\tau), \alpha(\tau)) A^{(i_0)} a_{i_0} P(\alpha_0 = a_{i_0}) d\tau. \tag{3.44}$$

where in the last line we use that for $lm_\mu \leq k \leq lm_\mu + m_\mu$ since $\varepsilon = \mu^{1+\Delta}$, $\varepsilon lm_\mu + m_\mu \leq \mu^\Delta(t+s) + \delta_\mu \to 0$ as $\mu \to 0$, by Remark 2.3 we have that $(P^\varepsilon)^{k-lm_\mu} \to I$ and $(P^\varepsilon)^{lm_\mu} \to I$ as $\mu \to 0$. The rest follows as before. $\square$

Fast Markov Chain: $\varepsilon \gg \mu$

In the case $\varepsilon \gg \mu$, the limit is characterized by the stationary distribution ν associated with Q.

Theorem 3.7. *Let $\varepsilon = \mu^\gamma$ for $1/2 \leq \gamma < 1$ and assume A3.1 – A3.4. Then $\theta^\mu(\cdot) \xrightarrow{w} \theta(\cdot)$ such that $\theta(\cdot)$ is the solution to the ODE*

$$\frac{d}{dt}\theta(t) = \sum_{i=1}^{m_0} A^{(i)}\left(a_i - \theta(t)\right)\nu_i, \qquad \theta(0) = \theta_0 \tag{3.45}$$

Proof. Here, we exploit that as $\mu l m_\mu \to \tau$ we have $\varepsilon(k - lm_\mu) = \mu^\gamma(k - lm_\mu) \to \infty$. Thus by Remark 2.3 $\Xi_{ij}(\varepsilon l m_\mu, \varepsilon k) = \nu_j + O(\varepsilon + \lambda_1^{-(k-lm_\mu)})$ for some $0 < \lambda_1 < 1$. Thus in (3.41)

$$\lim_{\mu \to 0} \mathbb{E}h^{N,\mu} \sum_{l\delta_\mu=t}^{t+s} \sum_{j=1}^{m_0} \delta_\mu \nabla f'(\theta_{lm_\mu}^N, \alpha_{lm_\mu}) \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} A^{(j)} a_j E_{lm_\mu} \mathcal{I}_{\{\alpha_k=a_j\}} \tag{3.46}$$
$$= \mathbb{E}h^N \sum_{j=1}^{m_0} \int_t^{t+s} \nabla_\theta f(\theta(\tau), \alpha(\tau)) A^{(j)} a_j \nu_j d\tau.$$

The result follows. $\square$

3.3 Asymptotic Distributions

On-Line: $\varepsilon = O(\mu)$

Without loss of generality we take $\varepsilon = \mu$ in this section. Define the scaled error

$$u_n := \widetilde{\theta}_n/\sqrt{\mu} = (\alpha_n - \theta_n)/\sqrt{\mu}. \tag{3.47}$$

We note that

$$u_{n+1} = u_n - \sqrt{\mu}\varphi_n \mathrm{sgn}(\varphi_n'\widetilde{\theta}_n + e_n) + \frac{\alpha_{n+1} - \alpha_n}{\sqrt{\mu}}. \tag{3.48}$$

By Theorem 3.2 there is a $N_{\mu,\varepsilon} = N_\mu$ such that $E|\alpha_n - \theta_n|^2 = O(\mu)$ for $n \geq N_\mu$, with which we can show $\{u_n : n \geq N_\mu\}$ is tight. In addition, take N_μ large such that by (3.16), we have for $n \geq N_\mu$

$$\sum_{j=n}^{\infty} E_n(\alpha_{j+1} - \alpha_j) = O(\mu). \tag{3.49}$$

As in Chapter 2, we would then define the interpolated process as $u^\mu(t) = u_n$ for $t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$ to use the above bounds. However, we will omit the increment

shift (taking $N_\mu = 0$) to ease the burdensome notation as follows

$$u^\mu(t) \triangleq u_n \quad \text{for} \quad t \in [n\mu, n\mu + \mu). \tag{3.50}$$

As before, a truncation device may be employed to ensure the boundedness of the the scaled errors u_n. For notational simplicity, the boundedness will be assumed here. We begin by establishing the tightness of the sequence $\{u^\mu(\cdot)\}$ to ensure the existence of a weak limit.

Lemma 3.8. *The sequence* $\{u^\mu(\cdot)\}$ *is tight in* $D([0, \infty); \mathbb{R}^r)$.

Proof. Using (3.48) we see

$$u^\mu(t + s) - u^\mu(t) = -\sqrt{\mu} \sum_{k=t/\mu}^{(t+s)/\mu - 1} g_k + \frac{\alpha_{(t+s)/\mu} - \alpha_{t/\mu}}{\sqrt{\mu}}. \tag{3.51}$$

Using E_t^μ to denote the conditional expectation with respect to the σ-algebra $\mathcal{F}_t^\mu = \sigma\{u^\mu(\tau) : \tau \leq t\}$, we apply (3.49) to see that

$$E_t^\mu \left| u^\mu(t + s) - u^\mu(t) \right|^2 \leq K E_t^\mu \left| \sum_{k=t/\mu}^{(t+s)/\mu - 1} -\sqrt{\mu} g_k \right|^2 + O(\sqrt{\mu}) \tag{3.52}$$

Considering the first term, we observe

$$
\begin{aligned}
E_t^\mu \left| \sum_{k=t/\mu}^{(t+s)/\mu - 1} -\sqrt{\mu} g_k \right|^2 &= \sum_{i=1}^{m_0} E_t^\mu \mu \sum_{k=t/\mu}^{(t+s)/\mu - 1} \sum_{j=t/\mu}^{(t+s)/\mu - 1} g_k' g_j \mathcal{I}_{\{\alpha_k = a_i = \alpha_j\}} \\
&\leq \sum_{i=1}^{m_0} E_t^\mu \mu \sum_{k=t/\mu}^{(t+s)/\mu - 1} \sum_{j=t/\mu}^{(t+s)/\mu - 1} [A_k^{(i)} \widetilde{\theta}_k + o(\widetilde{\theta}_k)]'[A_j^{(i)} \widetilde{\theta}_j + o(\widetilde{\theta}_j)] \mathcal{I}_{\{\alpha_k = a_i\}} \mathcal{I}_{\{\alpha_j = a_i\}} \\
&\leq \sum_{i=1}^{m_0} E_t^\mu K \mu \left| \sum_{k=t/\mu}^{(t+s)/\mu - 1} (A_k^{(i)} - A^{(i)}) \widetilde{\theta}_k + o(\widetilde{\theta}_k) \right|^2 \mathcal{I}_{\{\alpha_k = a_i\}} \\
&\qquad\qquad + \sum_{i=1}^{m_0} E_t^\mu K \mu \left| \sum_{k=t/\mu}^{(t+s)/\mu - 1} A^{(i)} \widetilde{\theta}_k \right|^2 \mathcal{I}_{\{\alpha_k = a_i\}}.
\end{aligned}
\tag{3.53}
$$

By virtue of Theorem 3.2 we have $\mathbb{E}|\widetilde{\theta}_k|^2 = O(\mu)$ for $k \in [t/\mu, (t + s)/\mu)$ sufficiently large

(occurring when μ is sufficiently small). Thus in the last term of (3.53) we have

$$\mathbb{E}\sum_{i=1}^{m_0} E_t^\mu K\mu \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} A^{(i)}\widetilde{\theta}_k \right|^2 \mathcal{I}_{\{\alpha_k=a_i\}} \leq K\mu \sum_{i=1}^{m_0} \mathbb{E} \sum_{k=t/\mu}^{(t+s)/\mu-1} \left|\widetilde{\theta}_k\right|^2 \mathcal{I}_{\{\alpha_k=a_i\}} \leq O(\mu)s. \quad (3.54)$$

Applying the mixing inequality in A3.4 to the first term we have

$$\mathbb{E}\sum_{i=1}^{m_0} E_t^\mu K\mu \left| \sum_{k=t/\mu}^{(t+s)/\mu-1} (A_k^{(i)} - A^{(i)})\widetilde{\theta}_k + o(\widetilde{\theta}_k) \right|^2 \mathcal{I}_{\{\alpha_k=a_i\}}$$

$$\leq \mathbb{E}\sum_{i=1}^{m_0} E_t^\mu K\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} \sum_{j=t/\mu}^{(t+s)/\mu-1} [(A_k^{(i)} - A^{(i)})\widetilde{\theta}_k]'[(A_j^{(i)} - A^{(i)})\widetilde{\theta}_j]\mathcal{I}_{\{\alpha_k=a_i=\alpha_j\}}$$

$$+ K\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} E\left|\widetilde{\theta}_k\right|^2$$

$$\leq \mathbb{E}\sum_{i=1}^{m_0} K\mu \left[E_t^\mu \sum_{k=t/\mu}^{(t+s)/\mu-1} \left| (A_k^{(i)} - A^{(i)})\sqrt{\mu}u_k \right| \sum_{j\geq k} \left| (A_j^{(i)} - A^{(i)})\sqrt{\mu}u_j \right| \right]\mathcal{I}_{\{\alpha_k=a_i=\alpha_j\}} + O(\mu)s$$

$$\leq O(\mu)s$$

$$(3.55)$$

Thus for any $T < \infty$ and any $0 \leq t \leq T$,

$$\lim_{\delta\to 0} \limsup_{\mu\to 0} \left\{ \sup_{0\leq s\leq\delta} \mathbb{E}[E_t^\mu |u^\mu(s+t) - u^\mu(t)|^2] \right\} = 0$$

Applying the criterion [20, p.47] we have that $\{u^\mu(\cdot)\}$ is tight. $\qquad\square$

Observe that $g_k(a_i, i) = \varphi_k \text{sgn}(\varphi_k'[a_i - a_i] + e_k) = \varphi_k \text{sgn}(e_k)$, so that $g_k(a_i, i)$ is a mean zero mixing process by combination of A3.2 – A3.4. To proceed, we shall make use of the following variant of the well-known central limit theorem for mixing processes; see [3] or [7] for details.

We are now equipped to prove the main result.

Theorem 3.9. *If* $\varepsilon = O(\mu)$ *and under A3.1 – A3.4* $u^\mu(\cdot) \xrightarrow{w} u(\cdot)$ *such that*

$$du = -A^{(\alpha)}u\,dt - \widetilde{\Sigma}^{1/2}dw, \quad (3.56)$$

where $w(\cdot)$ is a standard Brownian motion and $\alpha = \alpha(\cdot)$ is the continuous-time Markov chain associated with Q.

Proof. As usual, extract a convergent subsequence of $u^\mu(\cdot)$ (still denoted by $u^\mu(\cdot)$) with limit $u(\cdot)$. We will show that for each $s, t > 0$, the limit process satisfies

$$u(t+s) - u(t) = \int_t^{t+s} -A^{(\alpha(\tau))} u(\tau) d\tau - \int_t^{t+s} \widetilde{\Sigma}^{1/2} dw(\tau) \tag{3.57}$$

Note from (3.51),

$$\begin{aligned}
u^\mu(t+s) - u^\mu(t) &= -\sqrt{\mu} \sum_{k=t/\mu}^{(t+s)/\mu-1} g_k + O(\sqrt{\mu}) \\
&= \sum_{i=1}^{m_0} [-\sqrt{\mu} \sum_{k=t/\mu}^{(t+s)/\mu-1} g_k] \mathcal{I}_{\{\alpha_k=a_i\}} + O(\sqrt{\mu}).
\end{aligned} \tag{3.58}$$

We define
$$g_k(i) \triangleq g_k I_{\{\alpha_k=a_i\}}, \quad \widetilde{g}_k(i) := E_k g_k(i), \quad \text{and}$$
$$\Delta_k(i) \triangleq [g_k(i) - g_k(a_i, i) - (\widetilde{g}_k(i) - \widetilde{g}_k(a_i, i))]$$

and expand on the (negative of the) inside of the sum indexed by i in (3.58) as

$$\begin{aligned}
&\sqrt{\mu} \sum_{k=t/\mu}^{(t+s)/\mu-1} g_k(i) \\
&= \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu} g_k(a_i, i) + \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu}[\widetilde{g}_k(i) - \widetilde{g}_k(a_i, i)] + \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu} \Delta_k(i) \\
&= \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu} \varpi_k + \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu[A_k^{(i)} u_k + o(|u_k|)] + \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu} \Delta_k(i).
\end{aligned} \tag{3.59}$$

using $\widetilde{g}_k(a_i, i) = o(\widetilde{\theta}_k) = o(\sqrt{\mu}|u_k|)$ by A3.3.

First, we show the last term in (3.59) is $o(1)$. Since $\Delta_k(i)$ is a martingale difference, we

have

$$
\mathbb{E}\Big|\sum_{k=t/\mu}^{(t+s)/\mu-1}\sqrt{\mu}\Delta_k(i)\Big|^2 = \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu\Big|\Delta_k(i)\Big|^2
$$

$$
= \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu\mathbb{E}[g_k(i)-g_k(a_i,i)]'[g_k(i)-g_k(a_i,i)] + \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu E[\widetilde{g}_k(i)-\widetilde{g}_k(a_i,i)]'[\widetilde{g}_k(i)-\widetilde{g}_k(a_i,i)].
$$

$$(3.60)$$

The boundedness of φ_k and u_k implies $\sqrt{\mu}\varphi_k' u_k \to 0$ in probability uniformly in k as $\mu \to 0$.

Hence, the first term in (3.60) has

$$
\sum_{k=t/\mu}^{(t+s)/\mu-1}\mu E[g_k(i)-g_k(a_i,i)]'[g_k(i)-g_k(a_i,i)]
$$

$$
= \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu E\varphi_k'\varphi_k[\mathrm{sgn}(\sqrt{\mu}\varphi_k' u_k + e_k) - \mathrm{sgn}(e_k)]^2 \to 0 \quad \text{as } \mu \to 0.
$$

$$(3.61)$$

Using A3.3 and A3.4, along with the boundedness of u_k, on the second term of (3.60) gives

$$
\sum_{k=t/\mu}^{(t+s)/\mu-1}\mu\mathbb{E}[\widetilde{g}_k(i)-\widetilde{g}_k(a_i,i)]'[\widetilde{g}_k(i)-\widetilde{g}_k(a_i,i)]
$$

$$
= \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu\mathbb{E}[\sqrt{\mu}A_k^{(i)}u_k + o(\sqrt{\mu}|u_k|)]'[\sqrt{\mu}A_k^{(i)}u_k + o(\sqrt{\mu}|u_k|)]
$$

$$
= \sum_{k=t/\mu}^{(t+s)/\mu-1}\mu^2\mathbb{E}\Big|(A_k^{(i)} - A^{(i)})u_k + A^{(i)}u_k + o(|u_k|)\Big|^2
$$

$$
\leq \mu^2 K \sum_{k=t/\mu}^{\infty}\phi(k-t/\mu) + \mu^2\sum_{k=t/\mu}^{(t+s)/\mu-1}K \to 0 \quad \text{as } \mu \to 0.
$$

$$(3.62)$$

Hence

$$
\mathbb{E}\sum_{k=t/\mu}^{(t+s)/\mu-1}\sqrt{\mu}\Delta_k(i) \to 0 \quad \text{as } \mu \to 0.
$$

Next, in the second term of (3.59) we have

$$\sum_{k=t/\mu}^{(t+s)/\mu-1} \mu[A_k^{(i)}u_k + o(|u_k|)]$$

$$= A^{(i)} \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu u_k + \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu(A_k^{(i)} - A^{(i)})u_k + \sum_{k=t/\mu}^{(t+s)/\mu-1} \mu o(|u_k|). \tag{3.63}$$

Similar to the previous section, choose a sequence m_μ such that $m_\mu \to \infty$ as $\mu \to 0$ but $\delta_\mu/\sqrt{\mu} = \sqrt{\mu}m_\mu \to 0$. Then

$$\sum_{k=t/\mu}^{(t+s)/\mu-1} \mu u_k = \sum_{l\delta_\mu=t}^{t+s} \delta_\mu u_{lm_\mu} + \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} [u_k - u_{lm_\mu}]. \tag{3.64}$$

Since for $lm_\mu \le k < lm_\mu + m_\mu$, $u_k - u_{lm_\mu} = O(\delta_\mu/\sqrt{\mu})$, so the second term above goes to 0 in probability, uniformly in t. Similarly, by A3.3,

$$\sum_{k=t/\mu}^{(t+s)/\mu-1} \mu(A_k^{(i)} - A^{(i)})u_k = \sum_{l\delta_\mu=t}^{t+s} \delta_\mu \frac{1}{m_\mu} \sum_{k=lm_\mu}^{lm_\mu+m_\mu-1} (A_k^{(i)} - A^{(i)})u_k \to 0. \tag{3.65}$$

Likewise, $\sum_{k=t/\mu}^{(t+s)/\mu-1} \mu o(|u_k|) \to 0$ in probability uniformly in t.

Hence, putting the above estimates together we obtain

$$u(t+s) - u(t) = \lim_{\mu\to 0} u^\mu(t+s) - u^\mu(t)$$

$$= \lim_{\mu\to 0} \sum_{i=1}^{m_0} \left[-A^i \sum_{l\delta_\mu=t}^{t+s} \delta_\mu u_{lm_\mu} \right] \mathcal{I}_{\{\alpha_k=a_i\}} - \sum_{k=t/\mu}^{(t+s)/\mu-1} \sqrt{\mu}\varpi_k \tag{3.66}$$

$$= -\int_t^{t+s} A^{\alpha(\tau)}u(\tau)d\tau - \int_t^{t+s} \widetilde{\Sigma}^{1/2}dw(\tau).$$

The Theorem follows.   □

Slower Markov Chain: $\varepsilon \ll \mu$

For the case $\varepsilon \ll \mu$, define

$$\alpha_* := \sum_{i=1}^{m_0} a_i \pi_{0,i}, \qquad v_n := \frac{\alpha_* - \theta_n}{\sqrt{\mu}}$$

$$v^\mu(t) := v_n \quad \text{for} \quad t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$$

$$A^{(*)} := \sum_{i=1}^{m_0} A^{(i)} \pi_{0,i}.$$

Then we have the following.

Theorem 3.10. *If $\varepsilon = \mu^{1+\eta}$ for some $0 < \eta \le 1$ and under A3.1 – A3.4 $v^\mu(\cdot) \xrightarrow{w} v(\cdot)$ such that*

$$dv = -A^{(*)} v dt - \widetilde{\Sigma}^{1/2} dw \tag{3.67}$$

where $w(\cdot)$ is a standard Brownian motion.

Faster Markov Chain: $\varepsilon \gg \mu$

For the case $\varepsilon \gg \mu$, define

$$\bar{\alpha} := \sum_{i=1}^{m_0} a_i \nu_i \qquad z_n := \frac{\bar{\alpha} - \theta_n}{\sqrt{\mu}}$$

$$z^\mu(t) := z_n \quad \text{for} \quad t \in [(n - N_\mu)\mu, (n - N_\mu)\mu + \mu)$$

$$\bar{A} := \sum_{i=1}^{m_0} A^{(i)} \nu_i.$$

Theorem 3.11. *If $\varepsilon = \mu^\gamma$ for some $1/2 \le \gamma < 1$ and under A3.1 – A3.4 $z^\mu(\cdot) \xrightarrow{w} z(\cdot)$ such that*

$$dz = -\bar{A} z dt - \widetilde{\Sigma}^{1/2} dw \tag{3.68}$$

where $w(\cdot)$ is a standard Brownian motion.

3.4   Numerical Experiments

Here we demonstrate the performance of the Sign-Error (SE) algorithms and compare it with the Sign-Regressor (SR) and Least Mean Squares (LMS) algorithms (see [28, 29], respectively). In contrast to the prior studies, the θ_n now is a Markov chain. For example, in [2] only slowly varying continuous signals were treated; the sign algorithms were proposed in [11] but only for a constant parameter; only slow Markov chains and adaptive stepsize algorithms were treated in [17]. Here, to highlight the multi-scale features, we treat three cases: $\varepsilon = (3/5)\mu$ ($\varepsilon = O(\mu)$); $\varepsilon = \mu^2$ (a slowly-varying Markov chain); and $\varepsilon = \sqrt{\mu}$ (a fast Markov chain). We fix the step size $\mu = .05$.

3.4.1   Matched Model

Here we observe an exactly matched model as in the problem formulation; that is, $y_n = \varphi'_n \alpha_n + e_n$. For ease of observation we take the regressors to be one-dimensional such that $\{\varphi_n\}$ and $\{e_n\}$ are i.i.d. $\mathcal{N}(0,1)$ and $\mathcal{N}(0,.25)$, respectively. For the Markov chain α_n we use state space $\mathcal{M} = \{-1, 0, 1\}$ with transition matrix $P^\varepsilon = I + \varepsilon Q$, where

$$
Q = \begin{bmatrix} -0.6 & 0.4 & 0.2 \\ 0.2 & -0.5 & 0.3 \\ 0.4 & 0.1 & -0.5 \end{bmatrix}
$$

is the generator of a continuous-time Markov chain whose stationary distribution is therefore $\nu = (1/3, 1/3, 1/3)$. Hence $\overline{\alpha} = \sum_{i=1}^{3} a_i \nu_i = 0$. We take the initial distribution for α_0 to be $(3/4, 1/8, 1/8)$. So $\alpha_* = \sum_{i=1}^{3} a_i P(\alpha_0 = a_i) = -0.625$. We proceed to observe 1000 iterations of the algorithm for the cases $\varepsilon = O(\mu)$ and $\varepsilon \gg \mu$, and $10,000$ iterations for the case $\varepsilon \ll \mu$ (in order to illustrate some variations of the parameter).

To observe the tracking behavior of the SE algorithm, in comparison to the SR and LMS
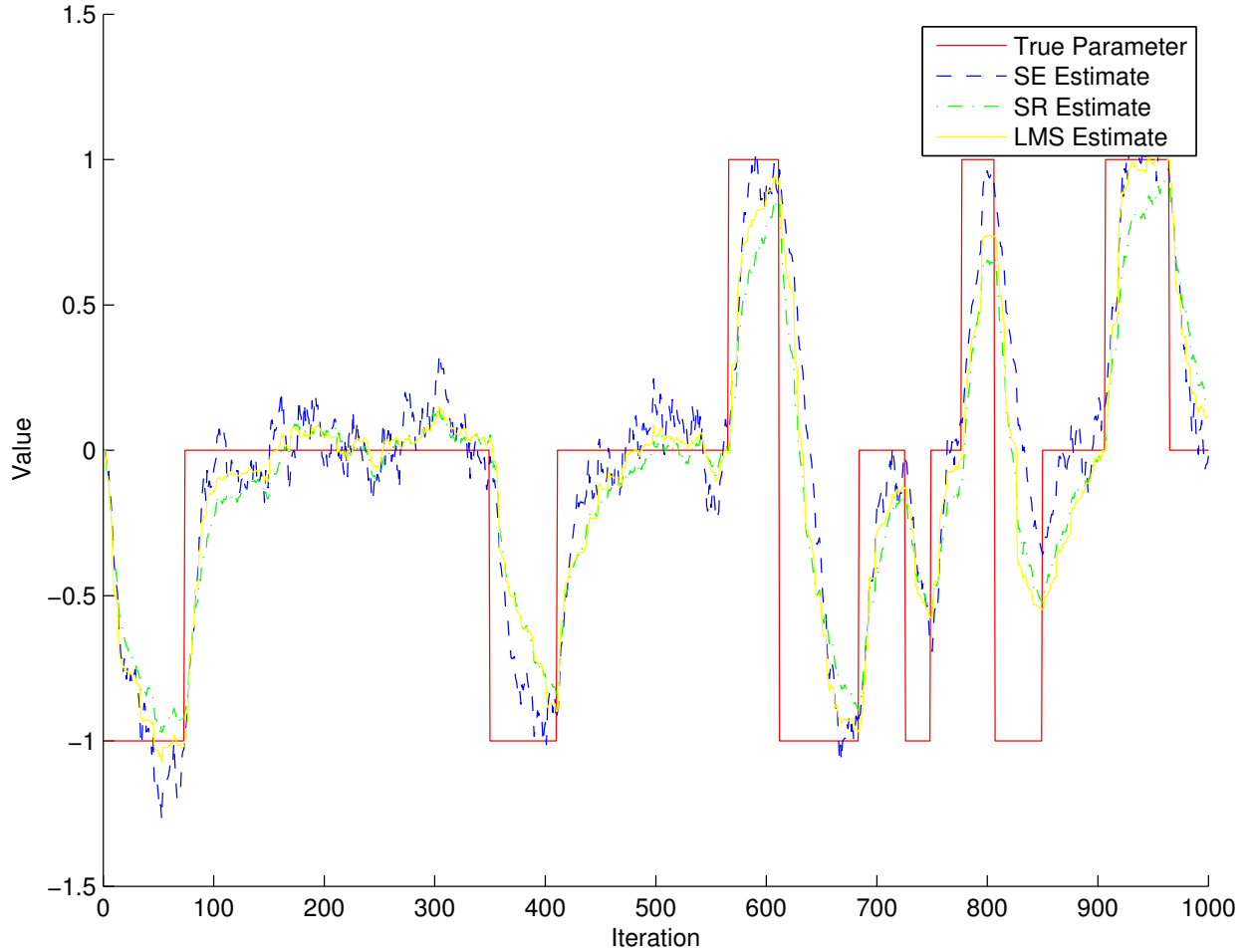
Figure 1: Adaptive filtering with On-Line Markovian parameter $\varepsilon = O(\mu)$

algorithms, we overlay the respective plots for each case. When $\varepsilon = O(\mu)$, the LMS and SR estimates tend to be approximately equal, while the SE estimates show more deviations from the other estimates. The SE algorithm responds to changes in the parameter more quickly, while the LMS and SR algorithms adhere to the parameter more closely while it is stationary. In the $\varepsilon \ll \mu$ case, we see this behavior repeated. While all three estimates track the parameter closely, the LMS and SR estimates deviate from the parameter less than the SE estimates between jumps of the parameter.

In the $\varepsilon \gg \mu$ case, none of the algorithms can track the parameter at each iterate very well. However, when we observe the scaled error against the stationary distribution of the Markov chain z_n, the diffusion behavior is displayed. Examining the cumulative average of
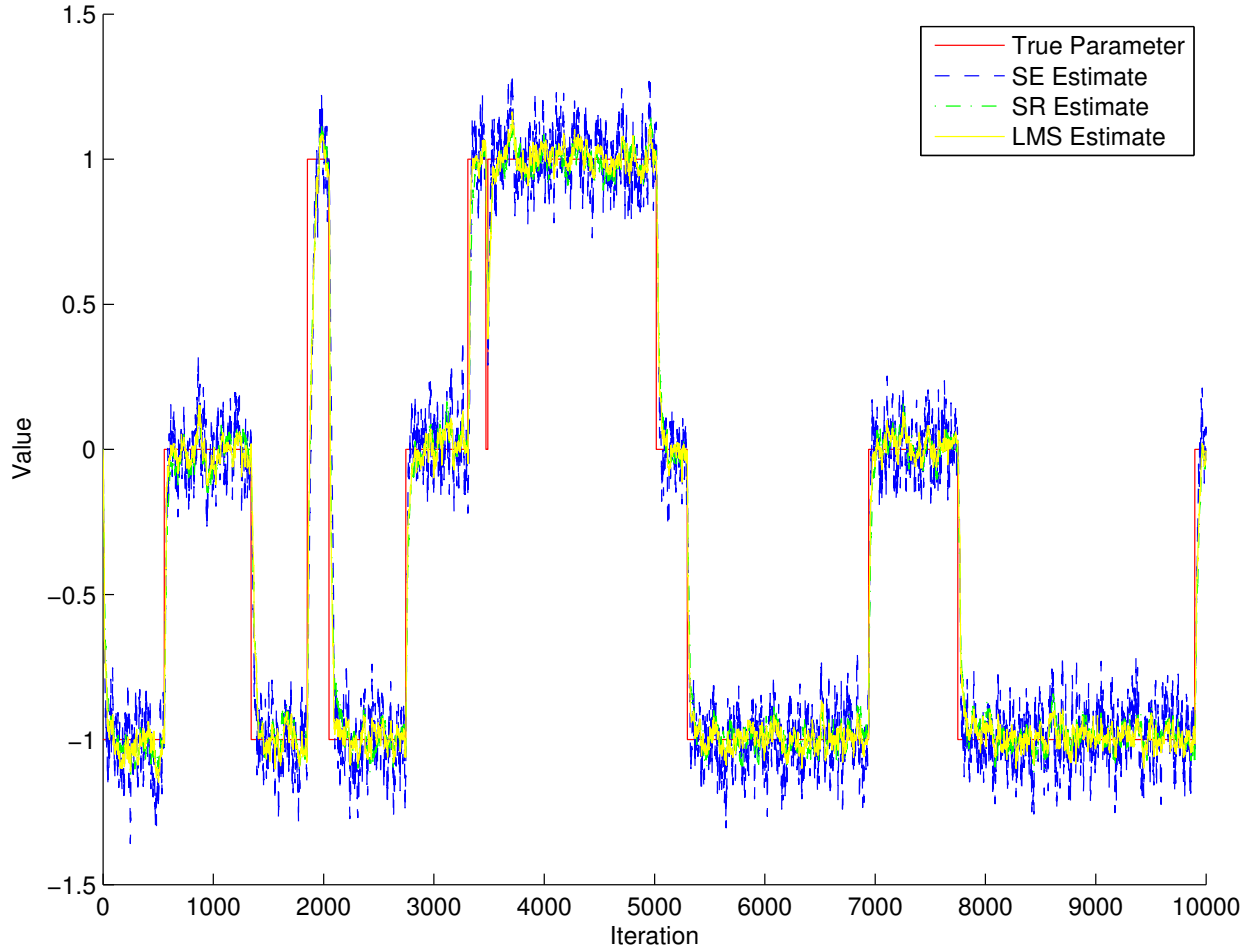
Figure 2: Adaptive filtering with Slower Markovian parameter $\varepsilon \ll O(\mu)$

the parameter and the estimates of the iterates, we note that the parameter average quickly converges to $\bar{\alpha}$. The LMS and SR estimate averages adhere closely to the parameter average, while the SE estimate average deviates slightly more.

3.4.2 Impact of Unmodeled Dynamics

We simulate the impact of unmodeled dynamics on the performance of the Sign-Error algorithm in the Markovian setting. We take the system given by $y_n = \varphi'_n \alpha_n + e_n = \check{\varphi}'_n \check{\alpha}_n + \widetilde{\varphi}'_n \widetilde{\alpha}_n + e_n$, where $\check{\varphi}_n$, $\check{\alpha}_n$ are the modeled parts of the regressors and parameters respectively, and $\widetilde{\varphi}_n$, $\widetilde{\alpha}_n$ are the unmodeled parts. We take φ_n i.i.d. 7-dimensional $\mathcal{N}(3,1)$, with modeled part $\check{\varphi}_n$ 4-dimensional, and errors $e_n \sim \mathcal{N}(0, .25)$ as before. For the Markov
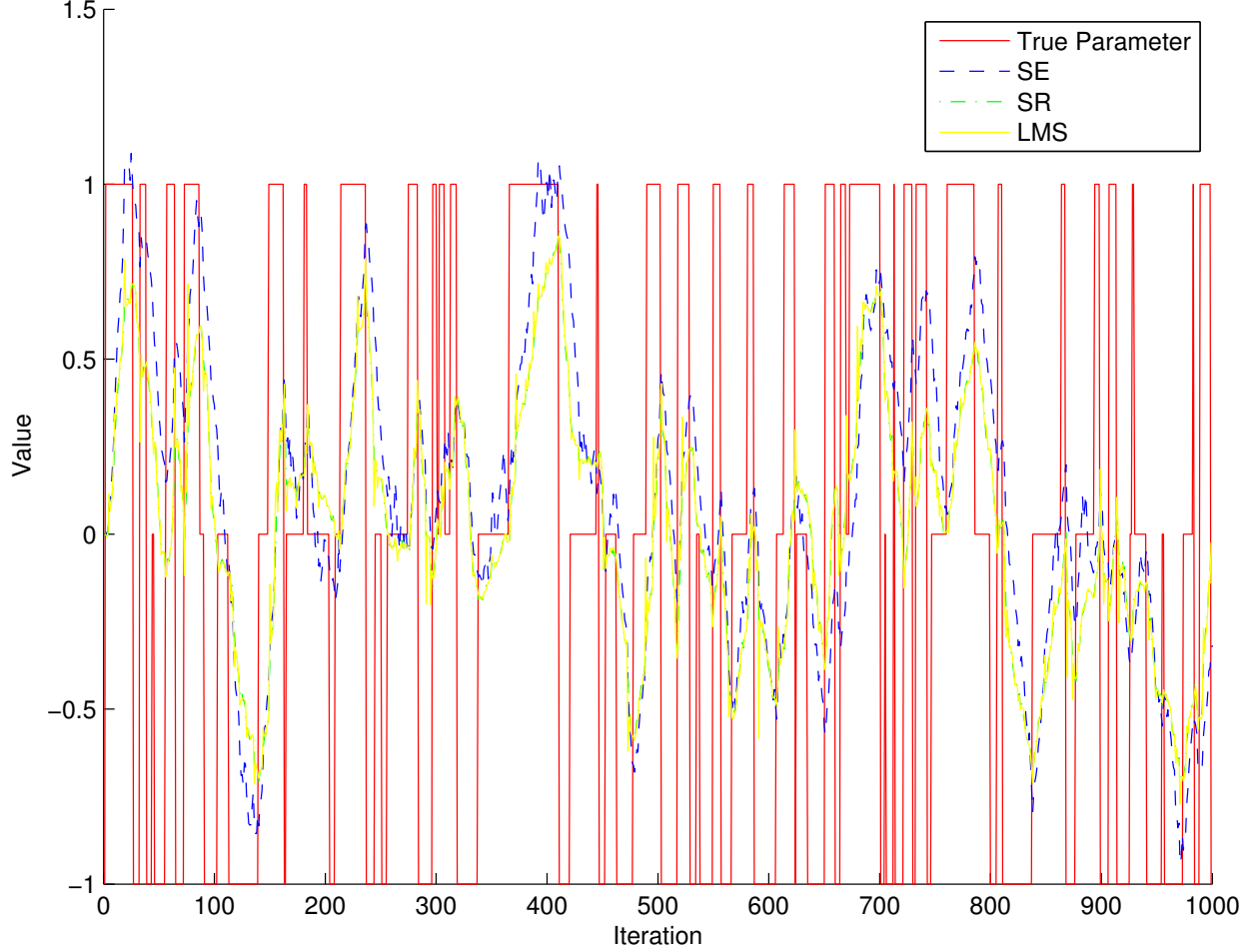
Figure 3: Adaptive Filtering with Fast Markovian parameter $\varepsilon \gg \mu$

chain we take state space $\mathcal{M} = \{-\rho, \mathbf{0}, \rho\}$ where $\rho = [1, 2^{-1}, \ldots, 2^{-6}] \in \mathbb{R}^7$. The transition matrix $P^\varepsilon = I + \varepsilon Q$ is as before, as well as the initial and stationary distributions.

We examine the SE algorithm for computing estimates θ_n, using the modeled part of the regressors $\check{\varphi}_n$ to track the modeled part of the parameter $\check{\alpha}_n$. More explicitly, $\theta_{n+1} = \theta_n + \check{\varphi}_n \operatorname{sgn}(y_n - \check{\varphi}'_n \theta_n) \in \mathbb{R}^4$. In Figure 6 we examine the norm difference between the modeled part of the parameter and the estimates $|\check{\alpha}_n - \theta_n|$ for $\varepsilon = O(\mu)$. In Figure 7 we see convergence in distribution by examining the average of the modeled part of the parameters $\bar{\check{\alpha}}_n$ up to time n and similarly for the estimates $\bar{\theta}_n$ (still for $\varepsilon = O(\mu)$). In Figure 8 we examine the difference of the time-averaged estimates $\bar{\theta}_n$ and (modeled) parameters $\bar{\check{\alpha}}_n$ from the modeled part of the stationary mean $\check{\bar{\alpha}}$ for the fast-varying case $\varepsilon \gg \mu$.
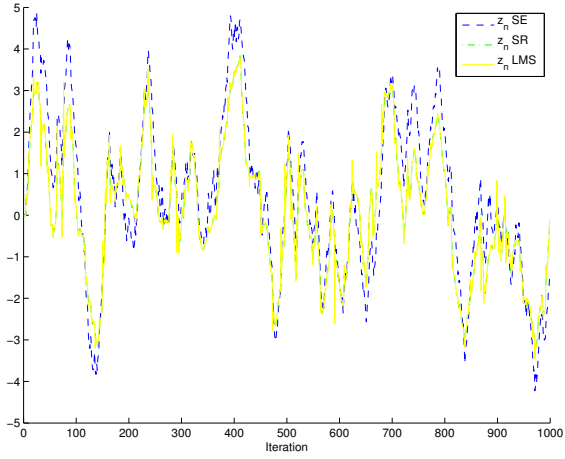
Figure 4: Scaled error z_n with fast varying Markov chain $\varepsilon \gg \mu$
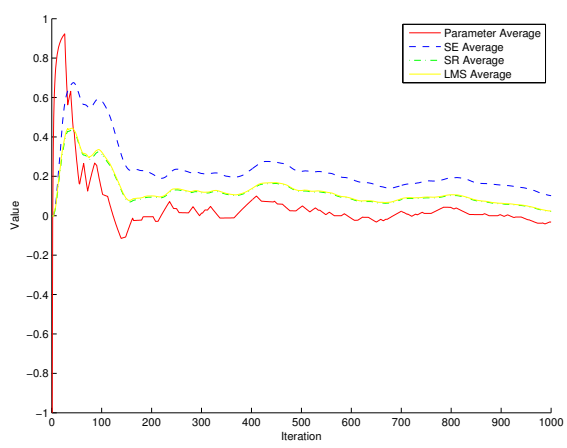
Figure 5: Average of parameter process and estimates over time with $\varepsilon \gg \mu$
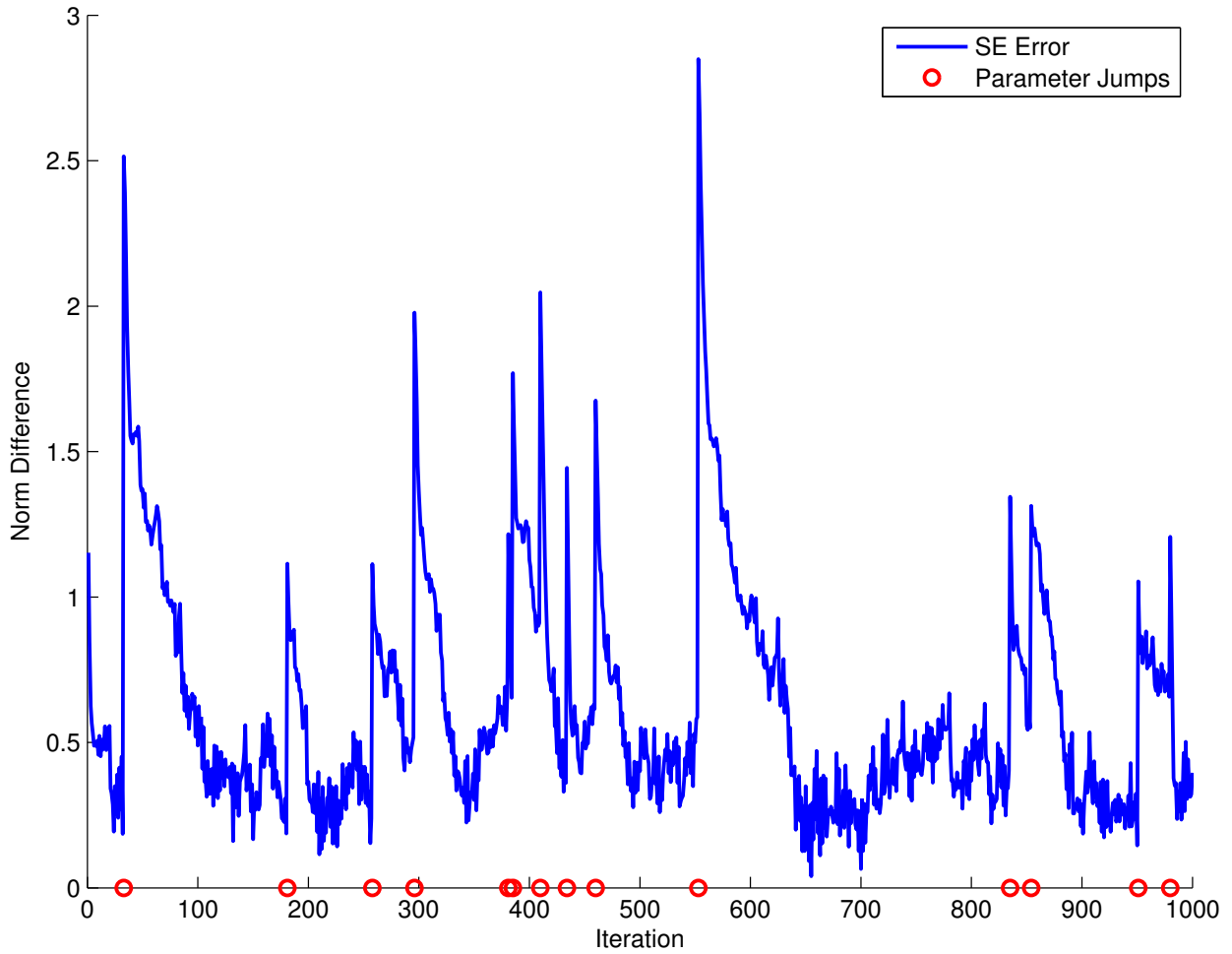

Figure 6: Difference of modeled parameter and estimate $|\check{\alpha}_n - \theta_n|$ for $\varepsilon = O(\mu)$.
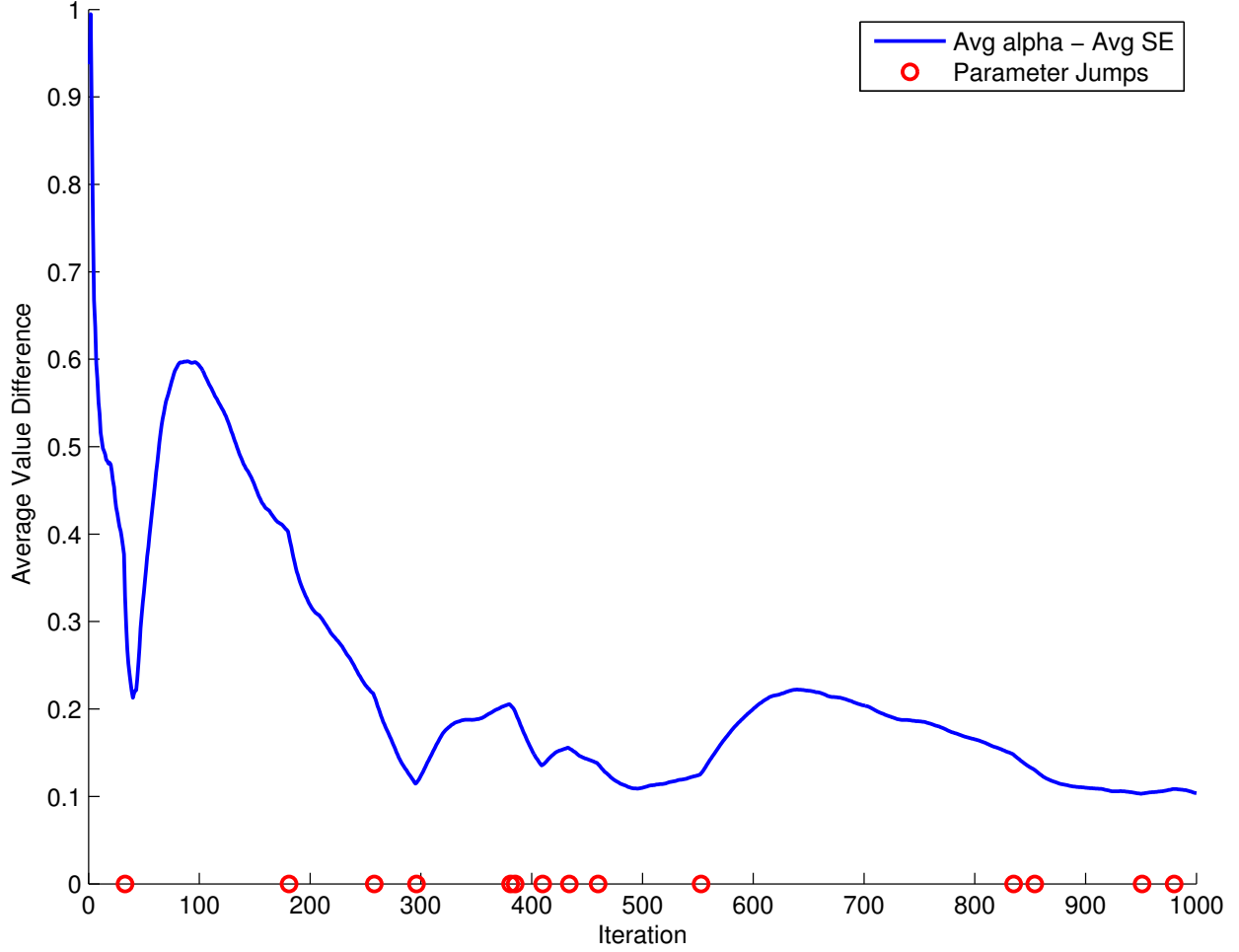
Figure 7: Average difference $|\bar{\bar{\alpha}}_n - \bar{\theta}_n|$ over time for $\varepsilon = O(\mu)$.

We note that with the influence of the unmodeled dynamics (of order 2^{-4}) against stochastic regressors of nonzero mean (order 3), there is a resulting bias in the residuals of order $3/2^4 \approx 0.2$. While the individual estimates vary slightly more in Figure 6, we see convergence within the unmodelled bias of order 0.2 in Figure 7. Similarly, in Figure 8 we observe convergence of both the estimates and parameters to the stationary mean within the 0.2 unmodeled bias.
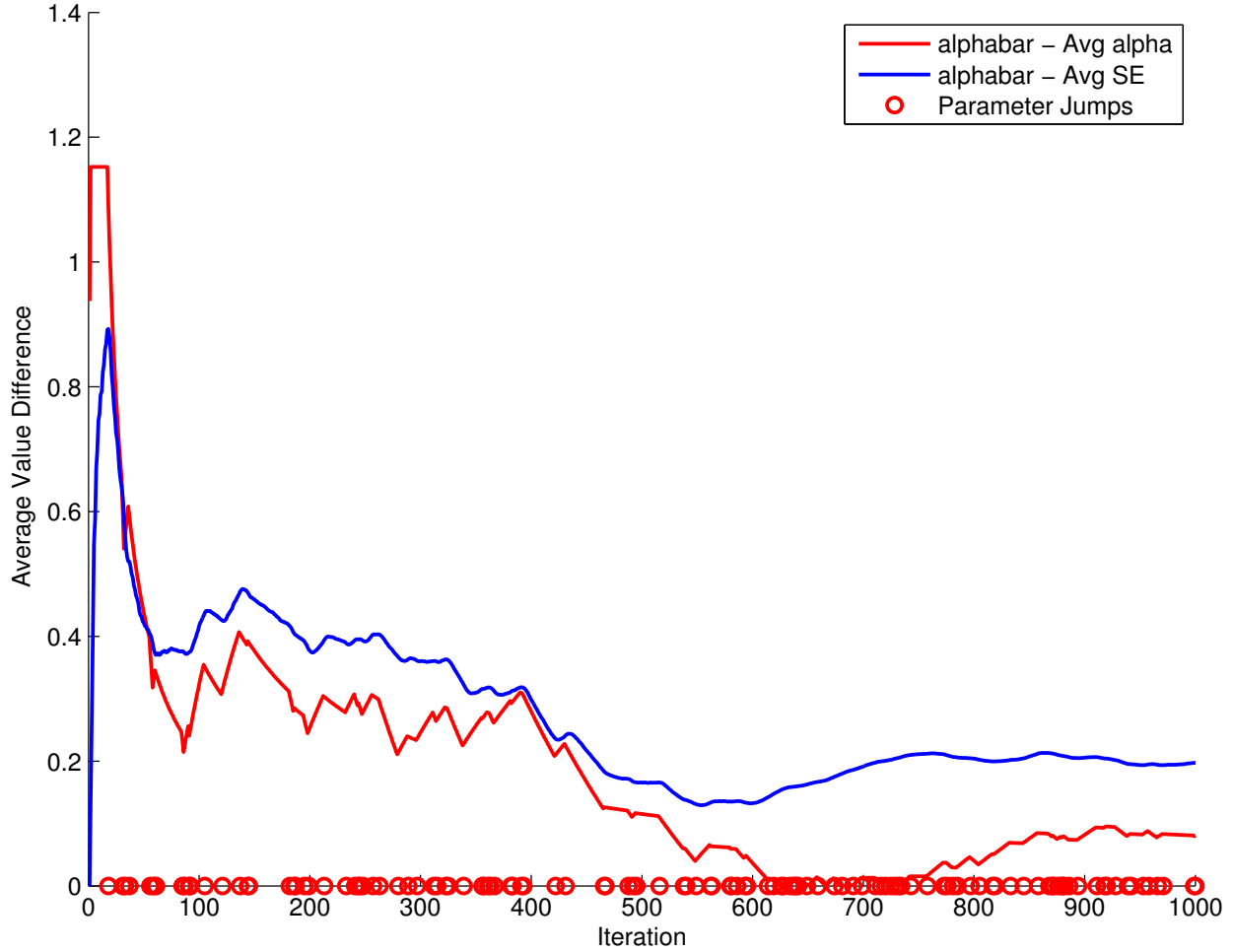
Figure 8: Average difference $|\bar{\bar{\alpha}}_n - \bar{\theta}_n|$ over time for $\varepsilon \gg \mu$.

4 Noise Attenuation with Unmodeled Dynamics

As demonstrated in the numerical experiments of the previous chapter, a mismatched model results in larger deviation from the limit. For example, oftentimes the output signal y_t is in fact a combination of *all* the previous input signals x_t; that is $y_t = \sum_{j=0}^{n} x_{t-j}\alpha_j$ known as an Infinite Impulse Response (IIR). However, in applications one must assume some finite model order n, that is $y_t = \sum_{j=0}^{\infty} x_{t-j}\alpha_j$ (Finite Impulse Response, FIR) and then applies an adaptive filtering algorithm such as in Chapters 2, 3 to estimate the underlying system parameters. The difference between the actual order of the system and the modeled order of the system introduces bias in the estimates due to the unmodeled dynamics. For tractability,

one assumes some bound ρ_n on the unmodeled bias which decays as the modeled order n increases. It is therefore desirable to employ filtering algorithms which are robust against the worst-case ρ_n for the unmodeled dynamics.

In what follows we develop such a robust filtering scheme for a regulation problem with a linear time-invariant (LTI) plant. The procedure is adaptive in nature, as it first estimates the noise for N steps, and then applies a control to attenuate the noise. We then analyze the impact of unmodeled dynamics and noise estimation errors by deriving error bounds, establishing a measure of robustness for the algorithm.

4.1 Motivation and Development

4.1.1 Linear Regulator Problem

Consider a regulation problem under the linear time invariant (LTI) plant P and controller F in Figure 9. The goal is to control the output x to follow the constant reference value x_r. However, the system output is influenced by stochastic disturbance d. Since the system is LTI, the output can be expressed in its transfer function form as

$$X(z) = \frac{F(z)P(z)}{1 + F(z)P(z)}X_r(z) + \frac{1}{1 + F(z)P(z)}D(z)$$
$$= U(z) + \frac{1}{1 + F(z)P(z)}D(z),$$

where the systems are represented by their z-transfer functions and the signals by their z transforms, and $U(z) = \frac{F(z)P(z)}{1+F(z)P(z)}X_r(z)$. x is measured. Denote $y_k = x_k - x_r$. Since x_k is measured and x_r is known, y_k is also a measured signal. Then

$$Y(z) = (U(z) - X_r(z)) + \frac{1}{1 + F(z)P(z)}D(z).$$

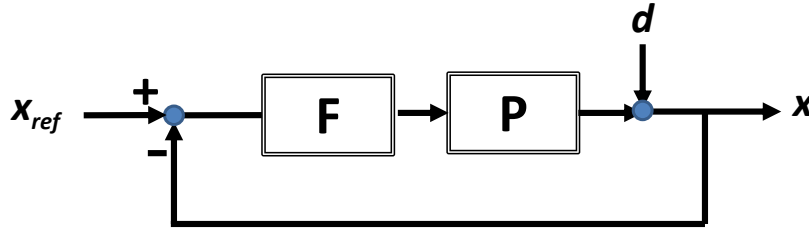We note the first term is deterministic and the second term is stochastic.

Figure 9: The original regulation problem

If the controller F is stabilizing and the system is at least of type 1 (including at least one integrator in the forward path), then the first term converges to zero exponentially fast. Since this is a very fast transient and our interest here is in noise rejection in a persistent sense, we will mandate a stabilizing controller in our design and then ignore this term in our analysis on noise attenuation. As a result, we will focus on

$$Y(z) = \frac{1}{1 + F(z)P(z)}D(z),$$

which can be represented by the diagram in Figure 10. Our goal is to attenuate the impact of the noise d on the output y. For simplicity, assume that P is an exponentially stable system.
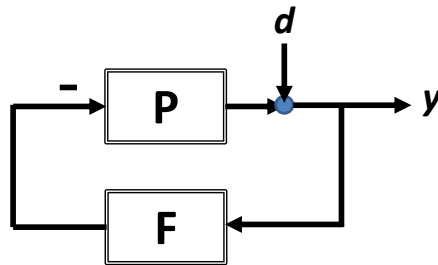

Figure 10: A basic feedback configuration for noise attenuation

Since the transfer function $P(z)$ is exponentially stable, we may represent it by a finite impulse response (FIR) filter $P_0(z)$ (the modeled part), plus an unmodeled dynamics δ:

$P(z) = P_0(z) + \delta(z)$. More precisely,

$$P(z) = p_0 + p_1 z^{-1} + \cdots + p_n z^{-n} + \delta(z) \tag{4.1}$$

where $\delta(z) = \displaystyle\sum_{j=n+1}^{\infty} p_j z^{-j}$ and $\displaystyle\sum_{j=n+1}^{\infty} |p_j| \leq \rho_n$. Due to exponential stability, $|\rho_n| \leq \kappa \lambda^n$ for some $\kappa > 0$ and $0 < \lambda < 1$, namely, it is an exponentially decaying function with respect to n. An immediate implication of this is that for a given required bound ρ on the modeling error, a model order n (model complexity) can be pre-determined such that $\rho_n \leq \rho$. In subsequent results, all bounds due to unmodeled dynamics should be interpreted as a function of model complexity n.

The following parametrization of stabilizing controllers is known as the Youla parametrization. In the special case of stable plants, it is called Q parametrization [9, 10].

Let $\mathbb{S}$ represent the space of exponentially stable systems. For internal stability, the closed-loop systems $\frac{1}{1+FP}$, $\frac{F}{1+FP}$, $\frac{P}{1+FP}$, $\frac{FP}{1+FP}$ must all be (exponentially) stable; that is, belong to $\mathbb{S}$. Denote:

$$Q = \frac{F}{1+FP} \in \mathbb{S}. \tag{4.2}$$

Since $P \in \mathbb{S}$, if $Q \in \mathbb{S}$, we have $PQ = \frac{FP}{1+FP} \in \mathbb{S}$, so $\frac{1}{1+FP} = 1 - \frac{FP}{1+FP} \in \mathbb{S}$, and hence $\frac{P}{1+FP} \in \mathbb{S}$.

Thus, the stability requirement is satisfied if we choose $Q \in \mathbb{S}$ and design $F = \frac{Q}{1-QP}$. This implies that F in Figure 10 can be implemented by using this Q parametrization, shown in Figure 11. Note that a positive feedback is used due to the presence of $1 - QP$ in the expression for F.

Let $Y(z)$ and $D(z)$ be the Laplace transforms of the output y and disturbance d respec-
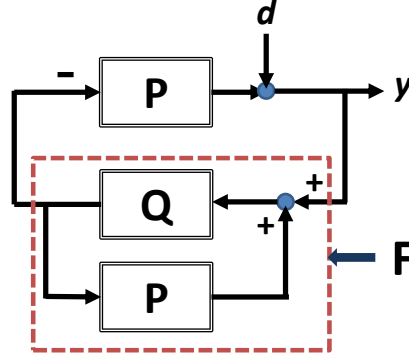
Figure 11: Feedback controller using the Q parameterization

tively. Then we have

$$Y(z) = \frac{1}{1+FP}D(z) = \left(1 - \frac{FP}{1+FP}\right)D(z) = (1 - QP)\,D(z).$$

Let $W(z) = P(z)D(z)$. Then $Y(z) = D(z) - QW(z)$. From

$$D(z) = d_0 + d_1 z^{-1} + \cdots; \quad W(z) = w_0 + w_1 z^{-1} + \cdots$$

we obtain the recursive representation

$$y_k = d_k - Q * w_k.$$

Suppose that Q is an FIR of order m. Then

$$
\begin{aligned}
y_k &= d_k - (q_0 w_k + q_1 w_{k-1} + \cdots + q_m w_{k-m}) \\
&= d_k - [w_k, w_{k-1}, \ldots, w_{k-m}][q_0, q_1, \ldots, q_m]' \\
&= d_k - \phi_k'\theta,
\end{aligned}
\tag{4.3}
$$

with $\phi'_k = [w_k, w_{k-1}, \ldots, w_{k-m}]$. Note that

$$
\begin{aligned}
w_k &= \sum_{j=0}^{\infty} p_j d_{k-j} = \sum_{j=0}^{n} p_j d_{k-j} + \sum_{j=n+1}^{\infty} p_j d_{k-j} \\
&= [d_k, d_{k-1}, \ldots, d_{k-n}]p + [d_{k-(n+1)}, \ldots]p^* \\
&= \psi'_k p + \widetilde{\psi}'_k p^*
\end{aligned}
\tag{4.4}
$$

where $p = [p_0, \ldots, p_n]'$ represents the modeled part of the plant and $p^* = [p_{n+1}, p_{n+2}, \ldots]'$ represents the unmodeled dynamics, and $\psi'_k = [d_k, d_{k-1}, \ldots, d_{k-n}]$, $\widetilde{\psi}'_k = [d_{k-(n+1)}, \ldots]$. We now introduce the following assumption to aid with the error analysis.

A 4.1. *(1) d_k is estimated by $\widehat{d}_k = d_k + e_k$. e_k is stationary, $\mathbb{E}e_k = 0$, $\mathbb{E}e_k^2 \leq \sigma^2 < \infty$.*

(2) The modeled part p is known. The unmodeled dynamics p^* has a uniform norm bound ρ_n.

Remark 4.1 (Signal Expansions). *Using Assumption A4.1, we can expand*

$$
\widehat{\psi}'_k = [\widehat{d}_k, \widehat{d}_{k-1}, \ldots, \widehat{d}_{k-n}] = [d_k + e_k, d_{k-1} + e_{k-1}, \ldots, d_{k-n} + e_{k-n}] = \psi'_k + \xi'_k
\tag{4.5}
$$

where $\xi'_k = [e_k, e_{k-1}, \ldots, e_{k-n}]$. Thus we can write

$$
w_k = \psi'_k p + \widetilde{\psi}'_k p^* = \widehat{\psi}'_k p - \xi'_k p + \widetilde{\psi}'_k p^* = \widehat{w}_k + \widetilde{\varepsilon}_k,
\tag{4.6}
$$

where

$$
\widehat{w}_k = \widehat{\psi}'_k p, \quad \widetilde{\varepsilon}_k = -\xi'_k p + \widetilde{\psi}'_k p^*.
\tag{4.7}
$$

As a result we have the decomposition

$$
\begin{aligned}
y_k &= d_k - [w_k, w_{k-1}, \ldots, w_{k-m}][q_0, q_1, \ldots, q_m]' \\
&= \widehat{d}_k - e_k - [\widehat{w}_k + \widetilde{\varepsilon}_k, \widehat{w}_{k-1} + \widetilde{\varepsilon}_{k-1}, \ldots, \widehat{w}_{k-m} + \widetilde{\varepsilon}_{k-m}][q_0, q_1, \ldots, q_m]' \\
&= \widehat{d}_k - e_k - \widehat{\phi}_k'\theta - \zeta_k'\theta,
\end{aligned}
$$

where $\widehat{\phi}_k' = [\widehat{w}_k, \widehat{w}_{k-1}, \ldots, \widehat{w}_{k-m}]$ and $\zeta_k' = [\widetilde{\varepsilon}_k, \widetilde{\varepsilon}_{k-1}, \ldots, \widetilde{\varepsilon}_{k-m}]$. For estimation, after N

observations the available regression data are

$$
\widehat{D}_N = \begin{bmatrix} \widehat{d}_1 \\ \vdots \\ \widehat{d}_N \end{bmatrix} ; \quad \widehat{\Phi}_N = \begin{bmatrix} \widehat{\phi}_1' \\ \vdots \\ \widehat{\phi}_N' \end{bmatrix}.
$$

In a nominal system based design procedure, the control parameter Q is then designed by

$$
\theta_N = \left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1} \widehat{\Phi}_N' \widehat{D}_N.
$$

If we define

$$
\Phi_N = \begin{bmatrix} \phi_1' \\ \vdots \\ \phi_N' \end{bmatrix} ; \quad \Xi_N = \begin{bmatrix} \zeta_1' \\ \vdots \\ \zeta_N' \end{bmatrix} ; \quad E_N = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}
$$

then

$$
\Phi_N = \widehat{\Phi}_N + \Xi_N \tag{4.8}
$$

and

$$Y_N = \widehat{D}_N - E_N - \widehat{\Phi}_N\theta - \Xi_N\theta. \tag{4.9}$$

The above calculations and expansions will be used in the error analysis.

4.1.2 Two-Phase Signal Estimation and Noise Rejection

We now discuss the signal estimation aspect of our noise attenuation scheme. In keeping with the adaptive filtering theme of this work, we note that in the above the control shall be defined and measured by estimates of the disturbances d_k. However, even though y_k is measured, d_k is not usually directly available. We proceed to explain why certain signals can be approximately extracted for control design.

After a controller F is (successfully) designed and implemented, the output $y_k = x_k - x_r$ will be small due to the rejection of disturbance by the feedback system. In this case y_k will have (nearly) no information which can be utilized for the control design. We shall call this phase the "noise rejection phase".

Before such a control is implemented, suppose that the disturbance d_k is stationary and its power spectrum density is limited in certain frequency bands. Then there exists an open-loop causal and stable filter $H(z)$ such that $H(z)D(z) \approx 0$ (i.e. $H(z)$ is an annihilating filter for d_k). If such a filter is inserted into the feedback loop in Figure 10 for a period of time, shown in Figure 12, the plant output v_k will be $V(z) = \frac{FP}{1+FPH}HD \approx 0$. Therefore during this phase, the signal d_k can be estimated by $y_k \approx d_k$. Thus in the subsequent control design, we should use the available signal

$$y_k = \widehat{d}_k = d_k + e_k \tag{4.10}$$

where $\widehat{d}_k$ is an estimate of d_k with estimation error e_k. We shall call this phase the "signal estimation phase".
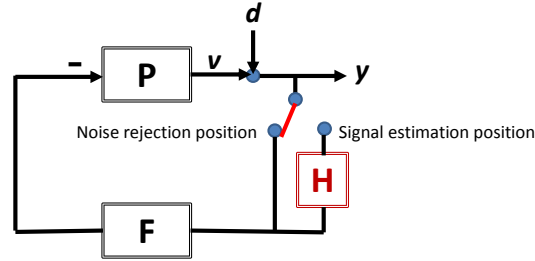

Figure 12: Modified configuration with annihilating filters for signal estimation

While the approach of using annihilating filters can potentially work for unstable plants, for stable plants a simple and general open-loop scheme works for the signal estimation phase. Suppose that $P(z)$ is stable with DC gain $K = P(1)$. By switching to the open-loop control illustrated in Figure 13, the actual output v_k of the plant is a deterministic (yet unknown) signal and converges exponentially to x_r. Thus the measured $y_k = x_k - x_r$ becomes d_k after an exponentially fast convergent transient. We note that this approach does not require any prior information on d_k. For this case, we also have $y_k = \widehat{d}_k = d_k + e_k$.
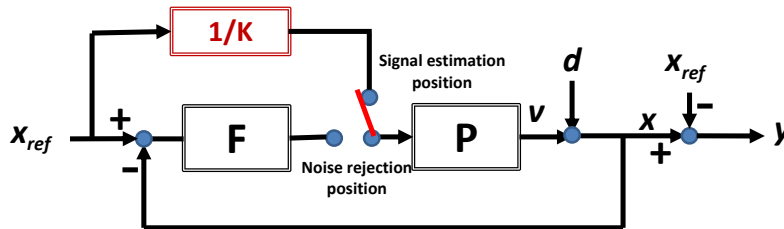

Figure 13: Use of open-loop control for signal estimation when the plant is stable

In this two-phase approach illustrated in Figure 14, control design is performed during the signal estimation phase. As a result, in the following algorithms, $\widehat{d}_k$ will be available in control design. The impact of signal estimation error e_k on noise rejection will be analyzed in Section 4.4.1.
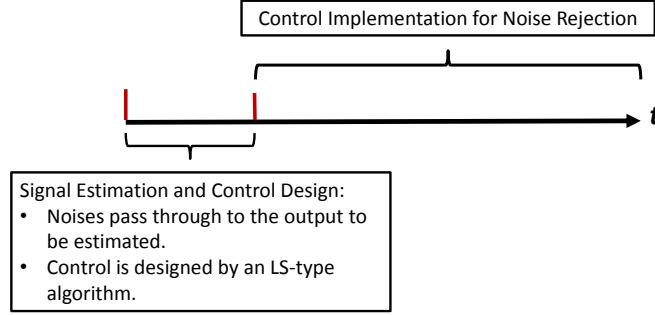
Figure 14: Two-phase design diagram

Example 4.2. For an example of this approach, we consider a plant with transfer function, before sampling, $\frac{1}{s+2}$. The DC gain of this plant is 0.5. As a result, the open-loop controller is $K = 2$. Suppose that the disturbance is $d_k = a_k \sin(200\tau k)$ where τ is the sampling interval and $\tau = 0.001$. a_k is i.i.d. and uniformly distributed in $[-5, 5]$. Now, in the first two seconds, we run this system open-loop. Then in $t \in [2, 10]$, we switch on the feedback controller which is a high gain feedback $F = 20000$. The trajectories of the disturbance d_k and the targeted y_k are shown in Figure 15.

4.2 Unmodeled Dynamics and Robust Noise Attenuation

In this section we analyze the impact of unmodeled dynamics and investigate suitable control design that can attenuate noise effects on the system output. Presently we shall focus only on the unmodeled dynamics, and so we take $e_k \equiv 0$ throughout this section. Consequently, we can simplify signals from Remark 4.1 to

$$\widehat{d}_k = d_k, \quad \widehat{\psi}'_k = \psi'_k = [d_k, \ldots, d_{k-n}],$$
$$\xi'_k = 0, \quad w_k = \psi'_k p + \widetilde{\psi}'_k p^*.$$
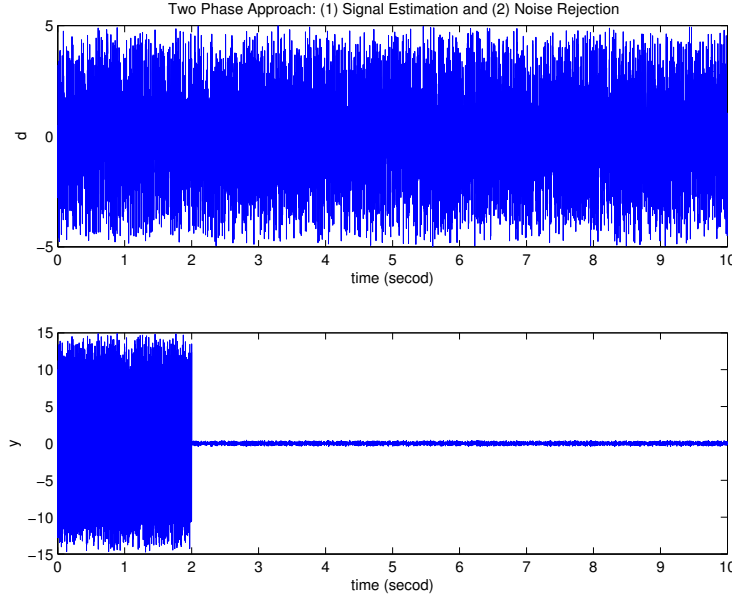
Figure 15: The two-phase design for Example 4.2

The observation equation (4.9) is simplified to

$$Y_N = D_N - (\widehat{\Phi}_N + \Xi_N)\theta,$$

with

$$\widehat{\Phi}_N = \begin{bmatrix} \widehat{\phi}'_1 \\ \vdots \\ \widehat{\phi}'_N \end{bmatrix}, \qquad \Xi_N = \begin{bmatrix} \zeta'_1 \\ \vdots \\ \zeta'_N \end{bmatrix}$$

and $\widehat{\phi}'_k = [\psi'_k p, \psi'_{k-1} p, \dots, \psi'_{k-m} p]$ and $\zeta'_k = [\widetilde{\psi}'_k p^*, \widetilde{\psi}'_{k-1} p^*, \dots, \widetilde{\psi}'_{k-m} p^*]$. We shall denote by $\Gamma \subset \mathbb{R}^{N \times m}$ the uncertainty set for the matrix Ξ_N which accomodates all possible unmodeled dynamics $p^* = \{p_j\}_{j=n+1}^{\infty}$ with $\sum_{j=n+1}^{\infty} |p_j| \le \rho_n$.

4.2.1 Nominal Design

Without signal estimation errors the disturbances d_k are directly measured and are available for the design phase. The nominal plant P_0 is known, and the nominal design is designed

without regard to the unmodeled dynamics. With the goal of minimizing the mean-square error $\min_{\theta_N} (D_N - \widehat{\Phi}_N \theta_N)'(D_N - \widehat{\Phi}_N \theta_N)$, the resulting control parameter θ becomes

$$\theta_N = \left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1} \widehat{\Phi}_N' D_N. \tag{4.11}$$

When using θ_N as in (4.11), we analyze performance by considering the residual of noise attenuation

$$
\begin{aligned}
\mu_N(\Xi_N, D_N) &\triangleq \frac{1}{N}(D_N - (\widehat{\Phi}_N + \Xi_N)\theta_N)'(D_N - (\widehat{\Phi}_N + \Xi_N)\theta_N) \\
&= \frac{1}{N}(D_N - (\widehat{\Phi}_N + \Xi_N)\left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1}\widehat{\Phi}_N' D_N)'(D_N - (\widehat{\Phi}_N + \Xi_N)\left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1}\widehat{\Phi}_N' D_N) \\
&= \frac{1}{N}D_N'(I - (\widehat{\Phi}_N + \Xi_N)\left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1}\widehat{\Phi}_N')'(I - (\widehat{\Phi}_N + \Xi_N)\left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1}\widehat{\Phi}_N')D_N \\
&= \frac{1}{N}D_N'\Pi'(D_N, \Xi_N)\Pi(D_N, \Xi_N)D_N
\end{aligned}
\tag{4.12}
$$

where

$$\Pi(D_N, \Xi_N) = I - (\widehat{\Phi}_N + \Xi_N)\left(\widehat{\Phi}_N' \widehat{\Phi}_N\right)^{-1}\widehat{\Phi}_N'$$

whose dependence on D_N stems from the fact that $\widehat{\Phi}_N$ depends on D_N. Define the worst-case performance as

$$\mu_N(D_N) \triangleq \max_{\Xi_N \in \Gamma} \mu(\Xi_N, D_N). \tag{4.13}$$

To proceed, we impose the following assumption on D_N. It is a sample-path version of disturbances' variances being bounded by σ^2.

A 4.2. *The N-sample path of the disturbances D_N satisfies*

$$D_N \in M_D \triangleq \{\|D_N/\sqrt{N}\|_2 \le \sigma^2\}.$$

To consider the worst-case performance of the disturbance attenuation, we first normalize the signal. Let $\|D_N/\sqrt{N}\|_2 = \lambda$ and define $v_N \overset{\Delta}{=} \frac{D_N/\sqrt{N}}{\lambda}$ so that $\|v_N\|_2 = 1$. For $D_N \in M_D$, $\lambda \leq \sigma^2$. Then we can obtain $\widehat{\Phi}_N(D_N) = \sqrt{N}\lambda\widehat{\Phi}_N(v_N)$. Denote $\sigma_{\min}$ as the smallest singular value of a matrix and

$$b_{\min} \overset{\Delta}{=} \min_{\|v_N\|_2=1} \sigma_{\min}(\widehat{\Phi}_N(v_N)).$$

Due to normalization, $b_{\min}$ is independent of the size of D_N. Also, denote

$$f(\rho_N) \overset{\Delta}{=} \max_{\Xi_N \in \Gamma} \frac{\|\Xi_N\|}{\sqrt{N}}$$

where ρ_N is the bound on unmodeled dynamics. We can now state the following theorem for the robust performance of the nominal design.

Theorem 4.3. *The worst-case disturbance attenuation performance is given by*

$$\mu \overset{\Delta}{=} \max_{D_N \in M_D} \mu_N(D_N) \leq \frac{f(\rho_N)}{b_{\min}} \tag{4.14}$$

Proof. Direct computation gives

$$\frac{1}{N}D'_N\Pi'(D_N,\Xi_N)\Pi(D_N,\Xi_N)D_N = v'_N(\lambda\Pi(D_N,\Xi_N))'(\lambda\Pi(D_N,\Xi_N))v_N,$$

in which

$$\begin{aligned}
\lambda\Pi(D_N,\Xi_N) &= \lambda(I - (\widehat{\Phi}_N(D_N) + \Xi_N)\left(\widehat{\Phi}'_N(D_N)\widehat{\Phi}_N(D_N)\right)^{-1}\widehat{\Phi}'_N(D_N)) \\
&= \lambda(I - (\sqrt{N}\lambda\widehat{\Phi}_N(v_N) + \Xi_N)\frac{1}{\sqrt{N}\lambda}\left(\widehat{\Phi}'_N(v_N)\widehat{\Phi}_N(v_N)\right)^{-1}\widehat{\Phi}'_N(v_N)) \\
&= \lambda(I - (\widehat{\Phi}_N(v_N) + \frac{\Xi_N}{\sqrt{N}\lambda})\left(\widehat{\Phi}'_N(v_N)\widehat{\Phi}_N(v_N)\right)^{-1}\widehat{\Phi}'_N(v_N)) \\
&\overset{\Delta}{=} \widehat{\Pi}(v_n,\Xi_N).
\end{aligned}$$

Additionally, we see

$$\mu = \max_{D_N \in M_D} \mu_N(D_N)$$

$$= \max_{\|v_N\|_2=1} \max_{\Xi_N \in \Gamma} v_N' \widehat{\Pi}(v_n, \Xi_N)' \widehat{\Pi}(v_n, \Xi_N) v_N,$$

so it follows that

$$\mu \leq \max_{\|v_N\|_2=1} \max_{\Xi_N \in \Gamma} \|\widehat{\Pi}(v_n, \Xi_N)\|$$

where $\|\cdot\|$ is the largest singular value. Using that

$$\widehat{\Pi}(v_N, \Xi_N)\widehat{\Phi}_N(v_N) = \lambda\left(I - (\widehat{\Phi}_N(v_N) + \frac{\Xi_N}{\sqrt{N}\lambda})\left(\widehat{\Phi}_N'(v_N)\widehat{\Phi}_N(v_N)\right)^{-1}\widehat{\Phi}_N'(v_N)\right)\widehat{\Phi}_N(v_N)$$

$$= \lambda\left(\widehat{\Phi}_N - (\widehat{\Phi}_N + \frac{\Xi_N}{\sqrt{N}\lambda})\right)$$

$$= -\frac{\Xi_N}{\sqrt{N}},$$

we have

$$\|\widehat{\Pi}(v_N, \Xi_N)\widehat{\Phi}_N(v_N)\| = \frac{\|\Xi_N\|}{\sqrt{N}} \leq \max_{\Xi_N \in \Gamma} \frac{\|\Xi_N\|}{\sqrt{N}} = f(\varepsilon_N).$$

With the above and

$$\|\widehat{\Pi}(v_N, \Xi_N)\widehat{\Phi}_N(v_N)\| \geq \|\widehat{\Pi}(v_N, \Xi_N)\|\sigma_{\min}(\widehat{\Phi}_N(v_N)) \geq \|\widehat{\Pi}(v_N, \Xi_N)\|b_{\min}$$

we obtain $\|\widehat{\Pi}(v_N, \Xi_N)\| \leq \frac{f(\varepsilon_N)}{b_{\min}}$ and hence the result

$$\mu \leq \frac{f(\varepsilon_N)}{b_{\min}}.$$

$\square$

4.2.2 Robust Design

In theory, when robustly attenuating noise for systems with unmodeled dynamics, one employs the performance index

$$\eta_N(D_N, \theta_N) \triangleq \frac{1}{N} \max_{\Xi_N \in \Gamma} (D_N - (\widehat{\Phi}_N + \Xi_N)\theta_N)'(D_N - (\widehat{\Phi}_N + \Xi_N)\theta_N), \qquad (4.15)$$

and seeks to find the optimal θ_N^* which attains

$$\eta_N(D_N) \triangleq \min_{\theta_N} \eta_N(D_N, \theta_N). \qquad (4.16)$$

The difference between the nominal design and robust design is that the former is a "max-min" design in which the design is done first; and the latter is a "min-max" design. One sees that

$$\eta_N(D_N) \leq \mu_N(D_N) \qquad (4.17)$$

indicating a potential performance improvement in the worst-case sense. It is well known that the "min-max" often leads to nonlinear and non-quadratic optimization problems and is usually more complicated. Often only numerical solutions are feasible, and to this end we proceed to introduce a gradient-descent numerical algorithm. Note that the gradient of $\eta_N(D_N, \theta_N)$ with respect to θ_N is

$$G(D_N, \theta_N) \triangleq \frac{\partial \eta_N(D_N, \theta_N)}{\partial \theta_N} = \frac{2}{N} \max_{\Xi_N \in \Gamma} (\widehat{\Phi}_N + \Xi_N)'(D_N - (\widehat{\Phi}_N + \Xi_N)\theta_N). \qquad (4.18)$$

Algorithm 4 (Two-Phase Algorithm). *The following algorithm searches for θ_N^* in two phases:*

- **Initial Value**.

The initial value θ^0 is given by the nominal design

$$\theta^0 = \left(\widehat{\Phi}'_N \widehat{\Phi}_N\right)^{-1} \widehat{\Phi}'_N D_N.$$

- **Iteration Steps**.

  For $k = 0, 1, 2, \ldots,$

  $$\theta^{k+1} = \theta^k - \beta_k \widehat{G}(D_N, \theta^k)$$

  where β_k is the step size at the kth iteration, $\widehat{G}(D_N, \theta^k)$ is an approximate gradient. Typically, these approximate values can be obtained by using Monte Carlo methods or grid calculation in place of the uncertainty set Γ.

4.3   Examples

We now use a simulation example to demonstrate performance on noise attenuation.

Example 4.4. The system to be controlled is a 7th order system $P(z) = p_0 + p_1 z^{-1} + \cdots + p_7 z^{-7}$. However, a lower-order model is used to represent this system: $P_0(z) = p_0 + p_1 z^{-1} + p_1 z^{-2} + p_3 z^{-3}$, leaving the higher-order terms as unmodeled dynamics. Hence, the modeled part has order $n = 3$ with 4 parameters, and the true values are $p_0 = 1$, $p_1 = 0.2$, $p_2 = 2$, and $p_3 = 0.5$. The unmodeled dynamics represent higher order terms which are excluded in the model, and in this example they are p_4, p_5, p_6, p_7. So, $p^* = [p_4, p_5, p_6, p_7]'$. We do not have information on the unmodeled dynamics, except for the bound $\rho = |p_4| + |p_5| + |p_6| + |p_7|$. In this example, we first use $\rho = 0.6$.

The noise sequence $\{d_k\}$ is i.i.d., uniformly distributed in $[-1, 1]$. As explained in the

previous sections, without estimation errors, d_k are known in our design process. The data length is $N = 1000$.

The uncertainty set from unmodeled dynamics is generated by the Monte Carlo method. We randomly generate 200 values of p^*, and then normalized them so that they all satisfy $|p_4| + |p_5| + |p_6| + |p_7| = 0.6$. The corresponding set of Ξ_N matrices is used as the uncertainty set Γ.

The controller has order $m = 20$, hence θ has 21 parameters. We consider the nominal design in this example. After generating the matrices D_N, Φ_N, we obtain

$$
\begin{aligned}
\theta_N &= \left(\widehat{\Phi}'_N \widehat{\Phi}_N\right)^{-1} \widehat{\Phi}'_N D_N \\
&= [0.0289, -0.1221, 0.4797, 0.0720, -0.2364, -0.0412, 0.1163, 0.0236, -0.0576, -0.0130, 0.0284, \\
&\quad 0.0075, -0.0143, -0.0039, 0.0072, 0.0020, -0.0035, -0.0013, 0.0016, 0.0009, -0.0010]'
\end{aligned}
$$

To evaluate performance on noise attenuation we use the noise-attenuation factor, defined as

$$
\gamma = \frac{\|Y_N\|_2/N}{\|D_N\|_2/N},
$$

where $\|D_N\|_2/N$ is the magnitude of the noise and $\|Y_N\|_2/N$ is the magnitude of the output. Thus $\gamma < 1$ indicates noise attenuation, and smaller γ corresponds to better noise-attenuation performance.

When there is no unmodeled dynamics ($\rho = 0$), the nominal design delivers a performance factor $\gamma = 0.0148$, which is an excellent 98.5% noise attenuation. However, when the unmodeled dynamics are introduced with $\rho = 0.6$, this factor is increased to $\gamma = 0.2943$ (70.1% noise reduction attenuation), a substantial loss of performance.

Figure 16 demonstrates noise attenuation performances. The top plot is the original un-attenuated noise, whose magnitude bound is 1. The second plot shows the noise attenuation performance of the controller when the system does not contain unmodeled dynamics. It is seen that the output values are around 0 and have much smaller magnitudes than the original noise, indicating substantial noise reduction. The third plot depicts the impact when the system contains unmodeled dynamics. By considering the worst case in the uncertainty set Γ, the noise reduction capability is significantly diminished when the nominally designed controller is used. To further illustrate this point, the fourth plot compares directly the performances between the matched-model system and the system with unmodeled dynamics. The first 500 points are the output when no unmodeled dynamics are involved, and the next 500 data points show impact of unmodeled dynamics. One sees the bias that results from the addition of the unmodeled component of the system. However, it should be noted that this is a worst-case study. There are some incidences in Γ under which the noise attenuation performance may be much better. This is the key issue of "robustness" of the controller which is assessed under the worst-case scenario.

Example 4.5. The impact of unmodeled dynamics on noise reduction performance is quite significant. To sustain acceptable noise reduction factors, one needs to use a well representative model so that the unmodeled dynamics are not too big. To illustrate such impact, we choose different sizes ρ for unmodeled dynamics for the same example as in Example 4.4 under the same simulation conditions. The resulting noise reduction factors and the corresponding noise reduction percentages are included in Table 1.
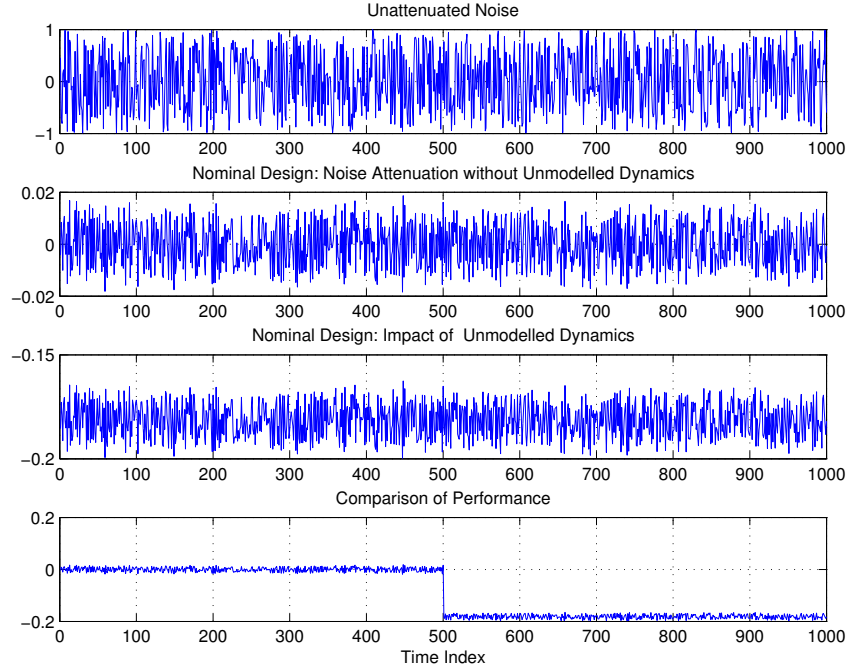
Figure 16: Noise attenuation under the nominal design.

Table 1: Impact of Unmodeled Dynamics

Size ρ of Unmodeled Dynamics	0.1	0.3	0.5	0.7	0.9
Reduction Factor	0.0570	0.1464	0.2512	0.3459	0.4493
Reduction Percentage	94.3%	85.4%	74.9%	65.4%	55.1%

4.4 Impact of Signal Estimation Errors

Lastly, we analyze impact of measurement errors by considering the difference of the system limit with only unmodeled dynamics from the limit with unmodeled dynamics and measurement error. We shall impose the following additional assumptions.

A 4.3. *The following conditions hold:*

1. *$\{d_k\}$ is a sequence of i.i.d. random variables satisfying $\mathbb{E}d_k = 0$ and $\mathbb{E}d_k^2 = \sigma_d^2 < \infty$. The fourth moment of d_k is finite: $\mathbb{E}d_k^4 < \infty$.*

2. *$\{d_k\}$ is estimated by $\widehat{d}_k = d_k + e_k$ such that $\{e_k\}$ is a sequence of independent and*

identically distributed (i.i.d.) random variables with $\mathbb{E}e_k = 0$ and $\mathbb{E}e_k^2 = \sigma_e^2 < \infty$. $\{e_k\}$

is independent of $\{d_k\}$.

3. *The modeled part p is known. The unmodeled dynamics p^* has a uniform norm bound*

ρ_n.

4.4.1 Limit with Measurement Errors

Let

$$
\begin{aligned}
\theta_N^e &= (\widehat{\Phi}_N' \widehat{\Phi}_N)^{-1} \widehat{\Phi}_N' \widehat{D}_N \\
&= ((\Phi_N - \Xi_N)'(\Phi_N - \Xi_N))^{-1}(\Phi_N' - \Xi_N')(D_N + E_N)
\end{aligned}
\tag{4.19}
$$

be the estimates from the design with both measurement errors and unmodeled dynamics. We begin by showing that for this nominal design the unmodeled dynamics are canceled out. This is done by separating the modeled and unmodeled components of Φ_N and Ξ_N as follows.

Recall that

$$
\Phi_N = \begin{bmatrix} \phi_1' \\ \phi_2' \\ \vdots \\ \phi_N' \end{bmatrix} = \begin{bmatrix} w_1 & w_0 & \cdots & w_{1-n} \\ w_2 & w_1 & \cdots & w_{2-n} \\ \vdots & \vdots & \vdots & \vdots \\ w_N & w_{N-1} & \cdots & w_{N-n} \end{bmatrix}.
$$

We separate the modeled and unmodeled parts of w_k by writing

$$
\begin{aligned}
w_k &= \psi_k' p + \widetilde{\psi}_k' p^* \\
&= \sum_{j=0}^{n} d_{k-j} p_j + \sum_{j=n+1}^{\infty} d_{k-j} p_j \\
&=: w_k^0 + \widetilde{w}_k,
\end{aligned}
$$

where $w_k^0 := \sum_{j=0}^{n} d_{k-j} p_j$ is a stationary, mean zero, strong mixing process [3] as d_k is i.i.d.

mean zero. Thus we may represent Φ_N by

$$\Phi_N = W_N^0 + \widetilde{W}_N$$

where W_N^0 and $\widetilde{W}_N$ are the $N \times (n+1)$ matrix collections of w_k^0 and $\widetilde{w}_k$ respectively. Also, we have

$$\Xi_N = \begin{bmatrix} \zeta_1' \\ \zeta_2' \\ \vdots \\ \zeta_N' \end{bmatrix} = \begin{bmatrix} \widetilde{\varepsilon}_1 & \widetilde{\varepsilon}_0 & \cdots & \widetilde{\varepsilon}_{1-n} \\ \widetilde{\varepsilon}_2 & \widetilde{\varepsilon}_1 & \cdots & \widetilde{\varepsilon}_{2-n} \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{\varepsilon}_N & \widetilde{\varepsilon}_{N-1} & \cdots & \widetilde{\varepsilon}_{N-n} \end{bmatrix}.$$

where

$$\widetilde{\varepsilon}_k = \widetilde{\psi}_k' p^* - \xi_k' p$$

$$= \sum_{j=n+1}^{\infty} d_{k-j} p_j - \sum_{j=0}^{n} e_{k-j} p_j$$

$$\triangleq \widetilde{w}_k - \varepsilon_k^0$$

Thus we have the decomposition

$$\Xi_N = \widetilde{W}_N - \Upsilon_N$$

where Υ_N is the $N \times (n+1)$ matrix of $\varepsilon_k^0 = \sum_{j=0}^{n} e_{k-j} p_j$, a stationary, mean zero, ergodic process. With this new notation, we have

$$\widehat{\Phi}_N = \Phi_N - \Xi_N = W_N^0 + \Upsilon_N \tag{4.20}$$

and so

$$\begin{aligned} \theta_N^e &= \left[\widehat{\Phi}_N' \widehat{\Phi}_N \right]^{-1} \widehat{\Phi}_N' \widehat{D}_N \\ &= \left[\frac{N}{N} (W_N^0 + \Upsilon_N)'(W_N^0 + \Upsilon_N) \right]^{-1} (W_N^0 + \Upsilon_N)'(D_N + E_N) \tag{4.21} \\ &= A_N \frac{1}{N} \left(W_N^{0\,'} D_N + W_N^{0\,'} E_N + \Upsilon_N' D_N + \Upsilon_N' E_N \right), \end{aligned}$$

where

$$A_N := \left[\frac{1}{N} \left(W_N^{0\,\prime} W_N^0 + \Upsilon_N{}' W_N^0 + W_N^{0\,\prime} \Upsilon_N + \Upsilon_N{}' \Upsilon_N \right) \right]^{-1}.$$ (4.22)

Write

$$P_n^0 = \left[\sum_{j=0}^{n-|l_2-l_1|} p_j p_{j+|l_2-l_1|} \right]_{l_1,l_2=0,1,\dots,n}.$$ (4.23)

Then we can formulate the limit of the estimate θ_N^e in terms of P_n^0 as follows.

Proposition 4.6. *Under Assumption 4.3, assuming P_n^0 is full rank, we have*

$$\theta_N^e = \left[\widehat{\Phi}_N' \widehat{\Phi}_N \right]^{-1} \widehat{\Phi}_N' \widehat{D}_N \xrightarrow{a.s.} \left[P_n^0 \right]^{-1} \begin{bmatrix} p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad as\ N \to \infty.$$ (4.24)

Proof. Working with the terms of A_N, we see that

$$\frac{1}{N} W_N^{0\,\prime} W_N^0 = \frac{1}{N} \sum_{k=1}^{N} \begin{bmatrix} w_k^0 w_k^0 & w_k^0 w_{k-1}^0 & \cdots & w_k^0 w_{k-n}^0 \\ w_{k-1}^0 w_k^0 & w_{k-1}^0 w_{k-1}^0 & \cdots & w_{k-1}^0 w_{k-n}^0 \\ \vdots & \vdots & \vdots & \vdots \\ w_{k-n}^0 w_k^0 & w_{k-n}^0 w_{k-1}^0 & \cdots & w_{k-n}^0 w_{k-n}^0 \end{bmatrix}$$ (4.25)

with

$$
\begin{aligned}
\mathbb{E} w^0_{k-l_1} w^0_{k-l_2} &= \mathbb{E} \sum_{j_1=0}^{n} \sum_{j_2=0}^{n} d_{k-l_1-j_1} d_{k-l_2-j_2} p_{j_1} p_{j_2} \mathcal{I}_{\{l_1+j_1=l_2+j_2\}} \\
&= \sigma_d^2 \sum_{j=0}^{n-|l_2-l_1|} p_j p_{j+|l_2-l_1|}.
\end{aligned}
\tag{4.26}
$$

We claim that the stationary process $\{w^0_{k-l_1} w^0_{k-l_2}\}_k$ has mean

$$
m \triangleq \mathbb{E} w^0_{k-l_1} w^0_{k-l_2} = \sigma_d^2 \sum_{j=0}^{n-|l_2-l_1|} p_j p_{j+|l_2-l_1|}
$$

and

$$
R(h) \triangleq \mathbb{E} \left\{ w^0_{k+h-l_1} w^0_{k+h-l_2} w^0_{k-l_1} w^0_{k-l_2} \right\} - m^2 \to 0.
$$

as $h \to \infty$. Examining the first term, we see

$$
\begin{aligned}
&\mathbb{E} \left\{ w^0_{k+h-l_1} w^0_{k+h-l_2} w^0_{k-l_1} w^0_{k-l_2} \right\} \\
&= \mathbb{E} \sum_{j_1,\ldots,j_4=0}^{n} d_{k+h-l_1-j_1} d_{k+h-l_1-j_2} d_{k-l_1-j_3} d_{k-l_2-j_4} p_{j_1} p_{j_2} p_{j_3} p_{j_4}.
\end{aligned}
$$

For $h > 2n$, $k+h-l_1-j_1 > k+h-l_2-j_2$ for $l, j \in \{0, \ldots, n\}$, so $d_{k+h-l_1-j_1}$ is independent

of $d_{k+h-l_2-j_2}$, and thus we can reduce the terms in the sum to

$$
\mathbb{E}\{w^0_{k+h-l_1}w^0_{k+h-l_2}w^0_{k-l_1}w^0_{k-l_2}\}
$$

$$
= \mathbb{E}\sum_{j_1,\ldots,j_4=0}^{n} d_{k+h-l_1-j_1}d_{k+h-l_1-j_2}d_{k-l_1-j_3}d_{k-l_2-j_4}p_{j_1}p_{j_2}p_{j_3}p_{j_4}\mathcal{I}_{\{l_1+j_1=l_2+j_2\}}
$$

$$
= \sum_{j_1=0}^{n-|l_2-l_1|}\sum_{j_3=0}^{n-|l_2-l_1|} \mathbb{E}[d^2_{k+h-j_1}]\mathbb{E}[d^2_{k+h-j_3}]p_{j_1}p_{j_1+|l_2-l_1|}p_{j_3}p_{j_3+|l_2-l_1|}
$$

$$
= \sigma_d^4\left[\sum_{j=0}^{n-|l_2-l_1|} p_j p_{j+|l_2-l_1|}\right]^2 = m^2
$$

(4.27)

Thus the covariance function $R(h) = 0$ for $h > 2n$. Moreover, $\sum_{h=0}^{N-1} R(h)/N \to 0$. As a result, with $X_k = w^0_{k-l_1}w^0_{k-l_2}$, $\overline{X}_N = \frac{1}{N}\sum_{k=1}^{N} X_k \xrightarrow{L^2} m$ as $N \to \infty$ by [16, Theorem 9.5.1]. Moreover, since $R(h) = 0$ for $h > 2n$, $\{X_k\}$ is a strong mixing process [16, p. 488]. By virtue of [16, Theorems 9.5.6], $\{X_k\}$ is strongly ergodic, and by [16, Theorems 9.5.5], $\overline{X}_N \to m$ a.s. Using the ergodicity obtained above and (4.23),

$$
\frac{1}{N}W_N^{0\prime}W_N^0 \xrightarrow{a.s.} \sigma_d^2 P_n^0 \qquad \text{as } N \to \infty.
$$

(4.28)

Similar arguments yield

$$
\frac{1}{N}\Upsilon_N'\Upsilon_N \xrightarrow{a.s.} \sigma_e^2 P_n^0 \quad \text{and} \quad \frac{1}{N}W_N^{0\prime}\Upsilon_N \xrightarrow{a.s.} 0 \quad \text{as } N \to \infty,
$$

and thus

$$
A_N \xrightarrow{a.s.} \left(\sigma_d^2 + \sigma_e^2\right)^{-1}\left[P_n^0\right]^{-1} \quad \text{as } N \to \infty.
$$

Examining the terms that A_N is applied to in (4.21),

$$\frac{1}{N}\left(W_N^{0\,\prime}D_N + W_N^{0\,\prime}E_N + \Upsilon_N' D_N + \Upsilon_N' E_N\right)$$

$$= \frac{1}{N}\sum_{k=1}^{N}\begin{bmatrix} w_k^0 d_k + w_k^0 e_k + \varepsilon_k^0 d_k + \varepsilon_k^0 e_k \\ w_{k-1}^0 d_k + w_{k-1}^0 e_k + \varepsilon_{k-1}^0 d_k + \varepsilon_{k-1}^0 e_k \\ \vdots \\ w_{k-n}^0 d_k + w_{k-n}^0 e_k + \varepsilon_{k-n}^0 d_k + \varepsilon_{k-n}^0 e_k \end{bmatrix} \qquad (4.29)$$

where

$$\mathbb{E}w_{k-l}^0 d_k = \mathbb{E}\sum_{j=0}^{n} d_{k-l-j} d_k = \sigma_d^2 p_0 \mathcal{I}_{\{l=0\}}$$

$$\mathbb{E}w_{k-l}^0 e_k = \mathbb{E}\sum_{j=0}^{n} d_{k-l-j} e_k = 0$$

$$\mathbb{E}\varepsilon_{k-l}^0 d_k = \mathbb{E}\sum_{j=0}^{n} e_{k-l-j} d_k = 0$$

$$\mathbb{E}\varepsilon_{k-l}^0 e_k = \mathbb{E}\sum_{j=0}^{n} e_{k-l-j} e_k = \sigma_e^2 p_0 \mathcal{I}_{\{l=0\}}.$$

Inspecting the covariance function for $X_k = w_{k-l}^0 d_k$ we see that d_{k+h} is independent of d_{k+h-j} for any $j > 0$, so that

$$\mathbb{E}w_{k+h-l}^0 d_{k+h} w_{k-l}^0 d_k$$

$$= \mathbb{E}\sum_{j_1=0}^{n}\sum_{j_2=0}^{n} d_{k+h-l-j_1} d_{k+h} d_{k+h-l_2-j_2} d_k p_{j_1} p_{j_2} \mathcal{I}_{\{l=0, j_1=0, j_2=0\}} \qquad (4.30)$$

$$= \mathbb{E}d_{k+h}^2 d_k^2 p_0^2 = \sigma_d^4 p_0^2 \mathcal{I}_{\{l=0\}} = \left[\mathbb{E}w_{k-l}^0 d_k\right]^2 \qquad \text{if } h > 0.$$

Thus $\frac{1}{N}\sum_{k=1}^{N} w_{k-l}^0 d_k \overset{a.s.}{\to} \mathbb{E}w_{k-l}^0 d_k$, and similarly for the other terms of (4.29). Hence we have

$$\frac{1}{N}\left(W_N^{0\,\prime}D_N + W_N^{0\,\prime}E_N + \Upsilon_N'D_N + \Upsilon_N'E_N\right) \overset{a.s.}{\to} (\sigma_d^2 + \sigma_e^2)\begin{bmatrix} p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.31}$$

Using (4.21), (4.22), (4.31), and the limits obtained thus far, we have

$$\left[\widehat{\Phi}_N'\widehat{\Phi}_N\right]^{-1}\widehat{\Phi}_N'\widehat{D}_N \overset{a.s.}{\to} (\sigma_d^2 + \sigma_e^2)^{-1}\left[P_n^0\right]^{-1}(\sigma_d^2 + \sigma_e^2)\begin{bmatrix} p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{4.32}$$

The proposition is thus established. □

4.4.2 Limit without Measurement Errors

Without measurement errors, the estimates are simplified to

$$\theta_N^0 = (\Phi_N'\Phi_N)^{-1}\Phi_N'D_N = B_N\left(\frac{1}{N}\Phi_N'D_N\right), \tag{4.33}$$

where $B_N = \left[\frac{1}{N}\Phi_N'\Phi_N\right]^{-1}$. Denote

$$P_n = \left[\sum_{j=0}^{\infty} p_j p_{j+|l_2-l_1|}\right]_{l_1,l_2=0,1,\dots,n}. \tag{4.34}$$

As before, we can formulate the limit of θ_N^e in terms of P_n as follows.

Proposition 4.7. *Under Assumption 4.3 and assuming P_n is full rank, we have*

$$\theta_N^0 = [\Phi_N'\Phi_N]^{-1}\Phi_N D_N \xrightarrow{a.s.} [P_n]^{-1} \begin{bmatrix} p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad as\ N \to \infty. \qquad (4.35)$$

Proof. We have that

$$\frac{1}{N}\Phi_N'\Phi_N = \frac{1}{N}\sum_{k=1}^N \begin{bmatrix} w_k w_k & \cdots & w_k w_{k-n} \\ \vdots & \vdots & \vdots \\ w_{k-n}w_k & \cdots & w_{k-n}w_{k-n} \end{bmatrix}, \qquad (4.36)$$

and observe

$$\begin{aligned}
\mathbb{E}w_{k-l_1}w_{k-l_2} &= \mathbb{E}\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} d_{k-l_1-j_1}d_{k-l_2-j_2}p_{j_1}p_{j_2} \\
&= \mathbb{E}\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} d_{k-l_1-j_1}d_{k-l_2-j_2}p_{j_1}p_{j_2}\mathcal{I}_{\{l_1+j_1=l_2+j_2\}} \qquad (4.37) \\
&= \sigma_d^2\sum_{j=0}^{\infty}p_j p_{j+|l_2-l_1|}
\end{aligned}$$

Using the definition (4.34), we have

$$\mathbb{E}\left[w_{k-l_1}w_{k-l_2}\right]_{l_1,l_2=0,\ldots,n} = \sigma_d^2 P_n. \qquad (4.38)$$

Establishing that the product sequences $\{w_{k-l_1}w_{k-l_2}\}_k$ are ergodic is more complicated due to the infinite sum involved in w_k. We show it for $X_k = w_k w_k$, with the shifted products

done in a similar manner. Here, $[\mathbb{E} w_k w_k]^2 = \left[\sum_{j=0}^{\infty} \sigma_d^2 p_j^2 \right] = \sigma_d^4 \sum_{j_1} \sum_{j_2} p_{j_1}^2 p_{j_2}^2$, and

$$\mathbb{E} w_{k+h} w_{k+h} w_k w_k = \sum_{j_1,\ldots,j_4=0}^{\infty} \mathbb{E} d_{k+h-j_1} d_{k+h-j_2} d_{k-j_3} d_{k-j_4} p_{j_1} p_{j_2} p_{j_3} p_{j_4}.$$

For the expectation of a term to be non-zero, every index of $d_\_$ must be paired with another.

Writing $\mathcal{A} = \{(j_1, j_2, j_3, j_4) : j_1 = j_2, j_3 = j_4\}$, $\mathcal{B} = \{j_1 = j_3 + h, j_2 = j_4 + h\}$, and

$\mathcal{C} = \{j_1 = j_4 + h, j_2 = j_3 + h\}$, we have that the non-zero terms are precisely $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$.

Furthermore, $\mathcal{A} \cap \mathcal{B} = \mathcal{A} \cap \mathcal{C} = \mathcal{B} \cap \mathcal{C} = \mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$, so $\sum_{\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}} = \sum_{\mathcal{A}} + \sum_{\mathcal{B}} + \sum_{\mathcal{C}} - 2 \sum_{\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}}$

Then we can express

$$
\begin{aligned}
&\mathbb{E} w_{k+h} w_{k+h} w_k w_k \\
&= \sum_{j_1=0}^{\infty} \sum_{j_3=0}^{\infty} \mathbb{E} d_{k-j_1}^2 d_{k-j_3}^2 p_{j_1}^2 p_{j_3}^2 + \sum_{j_3=0}^{\infty} \sum_{j_4=0}^{\infty} \mathbb{E} d_{k-j_3}^2 d_{k-j_4}^2 p_{j_3} p_{j_3+h} p_{j_4} p_{j_4+h} \\
&\quad + \sum_{j_3=0}^{\infty} \sum_{j_4=0}^{\infty} \mathbb{E} d_{k-j_3}^2 d_{k-j_4}^2 p_{j_3+h} p_{j_3} p_{j_4+h} p_{j_4} - 2 \sum_{j=0}^{\infty} \mathbb{E} d_{k-j}^4 p_j^2 p_{j+h}^2 \\
&= \sum_{j_1 \neq j_2+h} \sum_{j_2=0}^{\infty} \sigma_d^4 p_{j_1}^2 p_{j_2}^2 + \sum_{j=0}^{\infty} \mathbb{E} d^4 p_j^2 p_{j+h}^2 + 2 \sum_{j_1 \neq j_2} \sum_{j_2=0}^{\infty} \sigma_d^4 p_{j_1} p_{j_1+h} p_{j_2} p_{j_2+h} \\
&\quad + 2 \sum_{j=0}^{\infty} \mathbb{E} d^4 p_j^2 p_{j+h}^2 - 2 \sum_{j=0}^{\infty} \mathbb{E} d^4 p_j^2 p_{j+h}^2 \\
&= \sum_{j_1 \neq j_2+h} \sum_{j_2=0}^{\infty} \sigma_d^4 p_{j_1}^2 p_{j_2}^2 + \sum_{j=0}^{\infty} \mathbb{E} d^4 p_j^2 p_{j+h}^2 + 2 \sum_{j_1 \neq j_2} \sum_{j_2=0}^{\infty} \sigma_d^4 p_{j_1} p_{j_1+h} p_{j_2} p_{j_2+h}
\end{aligned}
\tag{4.39}
$$

Thus the covariance $R(h) = \mathbb{E}\left\{w_{k+h}w_{k+h}w_kw_k\right\} - \left[\mathbb{E}\left\{w_kw_k\right\}\right]^2$ satisfies

$$
\begin{aligned}
|R(h)| &= \Bigg| \sum_{j_1 \neq j_2 + h} \sum_{j_2 = 0}^{\infty} \sigma_d^4 p_{j_1}^2 p_{j_2}^2 + \sum_{j=0}^{\infty} \mathbb{E}d^4 p_j^2 p_{j+h}^2 \\
&\quad + 2 \sum_{j_1 \neq j_2} \sum_{j_2 = 0}^{\infty} \sigma_d^4 p_{j_1} p_{j_1+h} p_{j_2} p_{j_2+h} - \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \sigma_d^4 p_{j_1}^2 p_{j_2}^2 \Bigg| \\
&= \Bigg| -\sigma_d^4 \sum_{j_1 = j_2 + h} \sum_{j_2 = 0}^{\infty} p_{j_1}^2 p_{j_2}^2 + \mathbb{E}d^4 \sum_{j=0}^{\infty} p_j^2 p_{j+h}^2 + 2\sigma_d^4 \sum_{j_1 \neq j_2} \sum_{j_2 = 0}^{\infty} p_{j_1} p_{j_1+h} p_{j_2} p_{j_2+h} \Bigg| \quad (4.40) \\
&\leq |p_h| \sigma_d^4 \sum_{j=0}^{\infty} p_j^2 + |p_h| \mathbb{E}d^4 \sum_{j=0}^{\infty} p_j^2 + |p_h| 2\sigma_d^4 \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} p_{j_1} p_{j_2} \\
&\leq |p_h| K \to 0 \quad \text{as } h \to \infty.
\end{aligned}
$$

Thus the process $\{w_{k+h}w_{k+h}w_kw_k\}$ is strong mixing as well. Similar argument as in the derivation of (4.28) yields that $\frac{1}{N}\Phi'_N\Phi_N \overset{a.s.}{\to} \sigma_d^2 P_n$. Recall that P_n is full rank,

$$
B_N \overset{a.s.}{\to} \sigma_d^{-2}[P_n]^{-1} \qquad \text{as } N \to \infty. \tag{4.41}
$$

Similarly,

$$
\frac{1}{N}\Phi'_N D_N = \frac{1}{N}\sum_{k=1}^{N}
\begin{bmatrix}
w_k d_k \\
w_{k-1} d_k \\
\vdots \\
w_{k-n} d_k
\end{bmatrix} \tag{4.42}
$$

where

$$
\mathbb{E}w_{k-l}d_k = \mathbb{E}\sum_{j=0}^{\infty} d_{k-l-j} d_k p_j = \sigma_d^2 p_0 \mathcal{I}_{\{l=0\}} \tag{4.43}
$$

and the covariance function decays asymptotically in a manner similar to (4.40), so that

$$\frac{1}{N}\Phi'_N D_N \overset{a.s.}{\to} \begin{bmatrix} \sigma_d^2 p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{as } N \to \infty. \tag{4.44}$$

Finally, we have

$$\theta_N^0 = B_N \left(\frac{1}{N}\Phi'_N D_N \right) \overset{a.s.}{\to} \sigma_d^{-2} \{P_n\}^{-1} \begin{bmatrix} \sigma_d^2 p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

as $N \to \infty$, and the result follows.

$\square$

4.4.3 Difference of Estimates

Combining Propositions 4.6 and 4.7, we finally arrive at the following theorem. It gives the impact of measurement errors by characterizing the limit of the difference system with measurement error to the system without.

Theorem 4.8. *Under the assumptions of Propositions 4.6 and 4.7 and assuming that $P_n^0 - P_n$*

is invertible, we have

$$\theta_N^e - \theta_N^0 \overset{a.s.}{\to} \left[P_n^0 - P_n \right]^{-1} \begin{bmatrix} p_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.45}$$

where

$$- \left[P_n^0 - P_n \right]_{l_1, l_2} = \sum_{j=n-|l_2-l_1|+1}^{\infty} p_j p_{j+|l_2-l_1|}.$$

Defining

$$\rho_n^{(l)} \triangleq \sum_{j=n+1}^{\infty} p_{j-l} p_j \leq \sum_{j=n+1}^{\infty} |p_j| \leq \rho_n \tag{4.46}$$

for sufficiently large n, we see that

$$\left[P_n^0 - P_n \right]^{-1} = - \begin{bmatrix} \rho_n^{(0)} & \rho_n^{(1)} & \rho_n^{(2)} & \cdots & \rho_n^{(n)} \\ \rho_n^{(1)} & \rho_n^{(0)} & \rho_n^{(1)} & \cdots & \rho_n^{(n-1)} \\ \rho_n^{(2)} & \rho_n^{(1)} & \rho_n^{(0)} & \cdots & \rho_n^{(n-2)} \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \rho_n^{(n)} & \rho_n^{(n-1)} & \cdots & \rho_n^{(1)} & \rho_n^{(0)} \end{bmatrix}^{-1} \tag{4.47}$$

for $l = |l_2 - l_1| \in \{0, 1, \ldots, n\}$.

Example 4.9. We conduct a simulation study to display the limit of the estimate differences.

We take $\{d_k\} \sim \mathcal{N}(0, 1)$ and $\{e_k\} \sim \mathcal{N}(0, .1)$, both iid. The plant is a stable system with

IIR coefficients $p_k = (0.5)^k$ for $k = 0, 1, \ldots$. The model order is selected as $n = 10$. We then observe the estimates θ_N^e, θ_N^0 for $N = 10, 20, \ldots, 1010$ (100 updates). Thus $\rho_n = 2 - \sum_{k=0}^{10} p_k = (.5)^{10} \approx 9.8 \times 10^{-4}$. Figure 17 shows that $||\theta_N^e - \theta_N^0||$ quickly converges to $O(\rho_n)$.
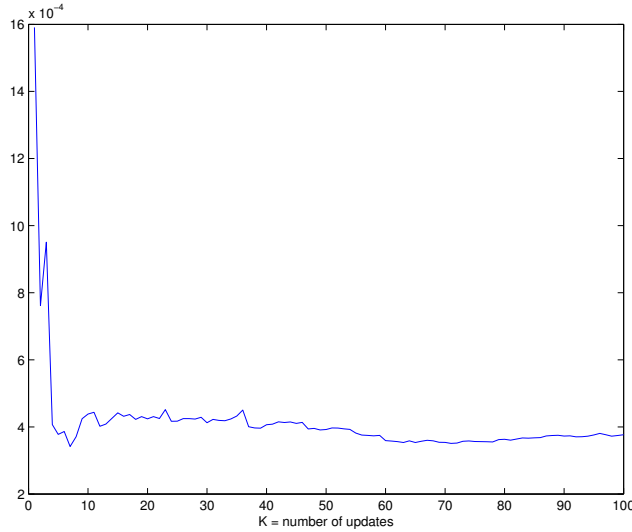

Figure 17:   Impact of est. errors given by $||\theta_N^e - \theta_N^0||$, $N = Kn = 10, \ldots, 1010$

5   Further Remarks

This dissertation has analyzed problems associated with adaptive filtering for identification and control of systems with switching Markovian dynamics, and of systems with unmodeled dynamics. For Markovian-switching systems, we used constant step-size algorithms to enable the estimates to persistently adapt to the changing dynamics of the underlying system. Error bounds and limit behavior was characterized by the relationship of the transition rate of the Markov chain to the adaptation rate of the estimates. For feedback systems with unmodeled dynamics, a two-phase algorithm was used to first estimate the noise characteristics and then attenuate the system to reduce the impact of the noise. Worst case performance bounds were obtained characterized by the magnitude of the unmodeled dynamics.

Naturally, a direction for future study would be analysis of systems with both Markovian-switching *and* unmodeled dynamics, as heuristically depicted in Figure 6. For example, consider a linear system given by

$$
\begin{aligned}
y_k &= \phi_k' \alpha_k + e_k \\
&= \check{\phi}_k' \check{\alpha}_k + \widetilde{\phi}_k' \widetilde{\alpha}_k + e_k
\end{aligned}
\tag{5.1}
$$

where $\phi_k' = [\check{\phi}_k', \widetilde{\phi}_k']$ is the input signal with modeled part $\check{\phi}_k \in \mathbb{R}^d$, $\alpha_k' = [\check{\alpha}_k', \widetilde{\alpha}_k']$ is the time-varying (Markovian) parameter with modeled part $\check{\alpha}_k \in \mathbb{R}^d$, and $e_k \in \mathbb{R}$ is the zero-mean noise at time k. Under the approach in Chapters 2, 3 one would use only the modeled component $\check{\phi}_k' \check{\alpha}_k$ to design the filter, and thus one expects to observe additional bias of magnitude ϵ_d resulting from the unmodeled component $\widetilde{\phi}_k' \widetilde{\alpha}_k$. This should result in an additional term of ϵ_d in the error bound, and an additional bias term in the infinitesimal limit.

In [30], such systems were studied with the Least Mean Squares algorithm estimates. It should be interesting to analyze performance of the Sign-Error algorithm with said systems. It may be that the direction-only scaling on the residuals employed in the SE algorithm could allow for better compensation of the bias in the unmodeled dynamics.

Additionally, we note the systems analyzed here were linear. If the underlying systems have nonlinear aspects, model mismatch will cause further bias in the estimates. A framework for addressing model mismatch, unmodeled dynamics, Markovian parameters, and stochastic noise was developed in [15]. Using said framework to analyze the robustness of the algorithms in this work is an important consideration to be addressed.

Finally, the step-size (adaptation rate) of the algorithms in Chapters 2, 3 were taken to be constant, and in Chapter 4 the step-sizes were left unspecified. As seen in the analysis, for a fixed step-size the performance of the algorithm depends on the underlying distribution of the Markov chain (the transition rate), and indeed can also be influenced by the possibly

time-varying bias of the unmodeled dynamics, model mismatch, and random noise. In such case, one may employ an additional adaptive algorithm to create a time-varying sequence of *stochastic* step-sizes $\{\mu_k\}$ which can respond to changes in the underlying dynamics to search for an optimal step-size. For example, in [15] the following adaptive step-size algorithm was presented for the Least Mean Squares algorithm

$$\theta_{k+1} = \theta_k + \mu_k \phi_k [y_k - \phi_k' \theta_k]$$

$$\mu_{k+1} = \Pi_{[\mu-,\mu+]} \left(\mu_k + c_1 [y_k - \varphi_k' V_k] \right) \tag{5.2}$$

$$V_{k+1} = V_k - \mu_k c_2 \phi_k \phi_k' V_k + \phi_k [y_k - \phi_k' \theta_k], \qquad V_0 = 0.$$

where c_1, c_2 are scaling constants, $[\mu-, \mu+]$ is a bounded set for the step-sizes μ, and $\Pi_{[\mu-,\mu+]}$ is the corresponding projection operator to ensure the iterates for μ_k remain in the feasible set. Similar considerations can be made for the SR and SE algorithms in a generalized framework.

REFERENCES

[1] A. Benveniste and G. Ruget. A measure of the tracking capability of recursive stochastic algorithms with constant gains. *IEEE Transactions on Automatic Control*, 27(3):639–649, 1982.

[2] Albert Benveniste, Michel Metivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer Publishing Company, Incorporated, 1st edition, 2012.

[3] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, September 2009.

[4] Han-Fu Chen and Lei Guo. *Identification and Stochastic Adaptive Control*. 1991.

[5] Han-Fu Chen and G. Yin. Asymptotic properties of sign algorithms for adaptive filtering. *IEEE Transactions on Automatic Control*, 48(9):1545–1556, September 2003.

[6] Sung Ho Cho and V.John Mathews. Tracking analysis of the sign algorithm in nonstationary environments. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(12):2046–2057, 1990. 2046.

[7] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

[8] Eweda Eweda. Optimum step size of sign algorithm for nonstationary adaptive filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1897–1901, 1990. 1897.

[9] B. Francis and G. Zames. On h-infinity optimal sensitivity theory for SISO feedback systems. *IEEE Transactions on Automatic Control*, 29(1):9–16, January 1984.

[10] Bruce A. Francis. *A Course in H Control Theory*. 1987.

[11] A. Gersho. Adaptive filtering with binary reinforcement. *IEEE Transactions on Information Theory*, 30(2):191–199, 1984.

[12] Lei Guo. Stability of recursive stochastic tracking algorithms. *SIAM Journal on Control and Optimization*, 32(5):1195–1225, September 1994.

[13] Araz Hashemi, Ben Fitzpatrick, Le Yi Wang, and George Yin. Robust noise attenuation under stochastic noises and worst-case unmodeled dynamics. *International Journal of Systems Science*, 2014.

[14] Araz Hashemi, G. Yin, and Le Yi Wang. Sign-error adaptive filtering algorithms for markovian parameters. *arXiv:1212.5185 [math]*, December 2012.

[15] Shaobai Kan, George Yin, and Le Yi Wang. Persistent tracking and identification of regime-switching systems with structural uncertainties: unmodeled dynamics, observation bias, and nonlinear model mismatch. *International Journal of Adaptive Control and Signal Processing*, 27(3):182198, 2013.

[16] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, 1975.

[17] V. Krishnamurthy, G. Yin, and S. Singh. Adaptive step-size algorithms for blind interference suppression in DS/CDMA systems. *IEEE Transactions on Signal Processing*, 49(1):190–201, January 2001.

[18] H. Kushner and A. Shwartz. Weak convergence and asymptotic properties of adaptive filters with constant gains. *IEEE Transactions on Information Theory*, 30(2):177–182, March 1984.

[19] Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*. Number 35 in Applications of mathematics. Springer-Verlag, New York, 2003.

[20] Harold Joseph Kushner. *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory.* MIT Press, January 1984.

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

[22] Ali H. Sayed. *Fundamentals of Adaptive Filtering.* Hoboken, N.J. : J. Wiley & Sons : IEEE Press, 2003.

[23] Le Yi Wang, G. George Yin, Ji-Feng Zhang, and Yanlong Zhao. *System Identification with Quantized Observations.* Springer, May 2010.

[24] S. Widrow and S.D. Stearns. *Adaptive Signal Processing.* Prentice-Hall, 1985.

[25] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series.* Wiley, New York, 1949.

[26] G. Yin. Adaptive filtering with averaging. *Adaptive Control, Filtering and Signal Processing; IMA Volumes in Mathematics and Its Applications*, 79:375–396, 1995.

[27] G. Yin, Vikram Krishnamurthy, and Cristina Ion. Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization. *SIAM Journal on Optimization*, 14(4):1187–1215, January 2004.

[28] G. George Yin, Araz Hashemi, and Le Yi Wang. Sign-regressor adaptive filtering algorithms for markovian parameters. *Asian Journal of Control*, 2013.

[29] G. George Yin and Vikram Krishnamurthy. Least mean square algorithms with markov regime-switching limit. *IEEE Transactions on Automatic Control*, 50(5):577–593, May 2005.

[30] G. George Yin, Shaobai Kan, Le Yi Wang, and Cheng-Zhong Xu. Identification of systems with regime switching and unmodeled dynamics. *IEEE Transactions on Automatic Control*, 54(1):34, January 2009.

[31] G. George Yin and Qing Zhang. *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications.* Springer, 2005.

[32] G. George Yin and Chao Zhu. *Hybrid Switching Diffusions: Properties and Applications.* Springer, 2010.

[33] Gang Yin and Hanfu Chen. On asymptotic properties of a constant-step-size sign-error algorithm for adaptive filtering. *Science in China Series F Information Sciences*, 45(5):321–334, October 2002.

[34] G.G. Yin, V. Krishnamurthy, and C. Ion. Iterate-averaging sign algorithms for adaptive filtering with applications to blind multiuser detection. *IEEE Transactions on Information Theory*, 49(3):657–671, March 2003.

[35] D. Youla, J. Bongiorno, and H. Jabr. Modern wiener–hopf design of optimal controllers part i: The single-input-output case. *IEEE Transactions on Automatic Control*, 21(1):3–13, February 1976.

[36] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2):301–320, April 1981.

ABSTRACT

ADAPTIVE STOCHASTIC SYSTEMS: ESTIMATION, FILTERING, AND NOISE ATTENUATION

by

ARAZ HASHEMI

MAY 2014

Advisor: Dr. G. George Yin

Major: Mathematics (Applied)

Degree: Doctor of Philosophy

This dissertation investigates problems arising in identification and control of stochastic systems. When the parameters determining the underlying systems are unknown and/or time varying, estimation and adaptive filtering are invoked to to identify parameters or to track time-varying systems. We begin by considering linear systems whose coefficients evolve as a slowly-varying Markov Chain. We propose three families of constant step-size (or gain size) algorithms for estimating and tracking the coefficient parameter: Least-Mean Squares (LMS), Sign-Regressor (SR), and Sign-Error (SE) algorithms.

The analysis is carried out in a multi-scale framework considering the relative size of the gain (rate of adaptation) to the transition rate of the Markovian system parameter. Mean-square error bounds are established, and weak convergence methods are employed to show the convergence of suitably interpolated sequences of estimates to solutions of systems of ordinary and stochastic differential equations with regime switching.

Next we consider problems in noise attenuation in systems with unmodeled dynamics and stochastic signal errors. A robust two-phase design procedure is developed which first estimates the signal in a simplified form, and then applies a control to tune out the noise. Worst-case error bounds are derived in terms of the unmodeled dynamics and variances of the disturbance and measurement errors.

AUTOBIOGRAPHICAL STATEMENT

ARAZ HASHEMI

Education : Wayne State Universtiy, Detroit, MI

- Ph.D. in Mathematics (Applied). May 2014

- M.A. in Mathematical Statistics, December 2012

- M.A. in Mathematics, May 2009

- B.S. in Mathematics, May 2007

Publications

- **A. Hashemi**, Y. Cao, D. Casbeer, and G. Yin, "UAV Circumnavigation of an Unknown Target Without Location Information,Using Noisy Range-Only Measurements", to appear in *American Control Conference*, (2014) available at arxiv.org/abs/1309.6007.

- **A. Hashemi**, B. Fitzpatrick, L.Y. Wang, and G. Yin, "Robust Noise Attenuation under Stochastic Noises and Worst-Case Unmodeled Dynamics", to appear in *International Journal of Systems Science* (2014).

- G. Yin, **A. Hashemi**, and L.Y. Wang, "Sign-Regressor Adaptive Filtering Algorithms for Markovian Parameters", *Asian Journal of Control*, (2013).

- **A. Hashemi**, Y. Cao, D. Casbeer, and G. Yin, "UAV Circumnavigation Using Noisy Range-based Measurements without GPS Information", submitted to *Journal of Dynamic Systems, Measurement, and Control*, (2014)

- **A. Hashemi**, G. Yin, and L.Y. Wang, "Sign-Error Adaptive Filtering Algorithms for Markovian Parameters", in review for *IEEE Transactions on Signal Processing*, available at arxiv.org/abs/1212.5185.

- Z. P. VandeGriend, **A. Hashemi**, and M. Shkoukani, "Changing Trends in Facial Trauma Epidemiology", in review for *JAMA Archives of Facial Plastic Surgery*.

Awards

- Karl and Helen Folley Endowed Mathematics Scholarship, WSU Mathematics, April 2013

- Alfred Nelson Award for Outstanding Achievement in Masters Program, WSU Mathematics, April 2013

- Alfred Nelson Award, WSU Mathematics, April 2009

- Outstanding Undergraduate Award, WSU Mathematics, April 2007

- Phi Beta Kappa, April 2007

- Honor Graduate, Cum Laude April 2007

- National Merit Special Scholar, National Merit Scholarship Coorporation, 2003-2007

- Presidential Scholar, Wayne State University, 2003-2007