Adaptive hidden Markov model estimation and applications

by Jason Ford

Bachelor of Science
Bachelor of Engineering


A thesis submitted for the degree of Doctor of Philosophy of the Australian National University.

Department of Systems Engineering
Research School of Information Sciences and Engineering
The Australian National University
Statement of Originality

These doctoral studies were conducted under the guidance of Professor John B. Moore as supervisor, and Dr. Peter Bartlett and Dr. Lei Wei as advisors.

The work submitted in this thesis is the result of original research carried out by myself, in collaboration with others, while enrolled in the Department of Systems Engineering as a Doctor of Philosophy student. It has not been submitted for any other degree or award in any other university or educational institution.

Jason John Ford.

Acknowledgements

I am indebted to Professor John Moore for his guidance, advice and enthusiasm as my supervisor and for providing a stimulating academic environment in the Department of Systems Engineering. I also thank Dr. Peter Bartlett and Dr. Lei Wei for their valuable advice as my academic advisors. I thank Professor Brian Anderson and Dr. Subhrakanti Dey for their advice, guidance and for their comments about my seminars and work at various stages throughout my doctoral studies. I thank Marita Rendina, Catrina Waterson and James Ashton for their help, guidance and humour over the last couple of years.

I am very grateful to Professor Robert Elliott (University of Alberta) for the fruitful collaborations I had with him. In hindsight, I'm also grateful for the discussions with Dr. Vikram Krishnamurthy (University of Melbourne) regarding parameter estimation for hidden Markov models. I have good memories of the collaborations with two summer students, Emily Lane (Auckland University) and Sanjeewa Athuraliya (University of Melbourne), I had the pleasure to supervise during their visits to the Australian National University.

I thank Dean Ray Spear from the Graduate School of the Australian National University for advice I received at various stages of my doctoral studies.

I thank the Co-operative Research Centre of Robust and Adaptive Systems for its financial assistance, and the opportunity to work on industrial problems. In particular, I thank Allan Connolly, Dr. Peter Kootsookos, and Glenn Dickins within the Robust Control Project for their collaborate approach to research and the experience I gained from working with them. The PhD experience would have been much less educational without the centre and its end sorrows me.

Finally, I am grateful to my fellow students, in particular, Louis Shue, Llew Mason, and Gordon Sutton for sharing a room with me and all that that entails. I thank former roommates Andrew Lim, Dr. Wee Sun Lee, Dr. Jeremy Mason and Dr. Ari Partanen for keeping me in
I thank Andrew Lim in particular for always listening to my questions and my answers to my own questions. I thank Carl Wendt and Peter Dower for the spicy, motivating and fruitful discussions over those beers. I thank my family for their support over the many years of my life. Last, but not least, I thank Ruth.
Abstract

Initially introduced in the late 1960's and early 1970's, hidden Markov models (HMMs) have become increasingly popular in the last decade. The major reason for the increasing popularity of HMMs has been the richness of the model class and the power of the signal processing tools.

In this thesis we propose several algorithms for estimation of HMM parameters. Initially, we propose recursive prediction error algorithms for separately estimating the state values and the state transition probability matrix. Local convergence results and corresponding convergence rates are obtained via an ordinary differential equation (ODE) approach. Suboptimal extended least squares algorithms are also presented and convergence results are established in idealized situations. These algorithms exploit the discrete-valued nature of HMMs.

Following this, globally convergent parameter estimators for HMMs are presented. These estimators have parallels to the well known Baum-Welch EM algorithm for off-line estimation of HMM parameters. Almost sure convergence results and convergence rates results are established using martingale convergence results, the Kronecker lemma and an ODE approach.

This inspires the proposal of globally convergent parameter estimators for partially observed linear systems and hybrid linear systems. Almost sure convergence results are established using martingale convergence results, the Kronecker lemma and an ODE approach.

Finally, as a contribution towards applications, optimal HMM filters are developed for demodulation of differentially encoded transmission systems and a decision feedback equalizer is proposed.
Preface

This thesis is mainly concerned with parameter estimation for hidden Markov models (HMMs). This thesis is divided into 7 chapters. Chapter 1 presents an introduction to the thesis, Chapters 2-6 form the body of the thesis and can be read independently of each other, and Chapter 7 is the conclusion to the thesis. In body of the thesis, Chapters 2, 3 and 4 present parameter estimators for HMMs, Chapter 5 presents almost sure parameter estimators for partially observed linear and hybrid linear systems and Chapter 6 investigates the use of HMM filters for state estimation of differentially encoded messages.

Most of the work presented in this thesis has been published, accepted, or has been submitted for publication, either as academic papers in refereed journals or conference proceedings. Some of the material in the conference papers overlaps with the journal papers. A list of the papers is given.

Journal Papers:


Conference Papers:


# Contents

1 **Introduction**  
1.1 Background and Motivation .................................................. 2  
1.1.1 Hidden Markov Models .................................................... 3  
1.1.2 Linear Systems and Hybrid Linear Systems ............................... 4  
1.1.3 System Identification ..................................................... 4  
1.1.4 A Brief History of Hidden Markov Model Estimation ..................... 7  
1.2 Outline of Thesis Contributions ............................................. 8  
1.3 Review of Various Mathematical Tools ...................................... 10  
1.4 Definitions ................................................................. 17  
1.5 Delimitation of Scope and Key Assumptions .................................. 18  

2 **Estimation of the State Values**  ........................................... 19  
2.1 Introduction ................................................................. 19  
2.2 Problem Formulation .......................................................... 21  
2.2.1 HMM State Space Model .................................................. 21  
2.2.2 Information State Model .................................................. 22  
2.3 Least Squares and Extended Least Squares .................................. 24  
2.3.1 Least Squares ............................................................... 24  
2.3.2 Extended Least Squares ................................................... 28  
2.4 Recursive Prediction Algorithms ............................................ 32  
2.5 Implementation Considerations and Simulations Results ................... 37  
2.5.1 Implementation Considerations ......................................... 37  
2.5.2 Simulations ............................................................... 39  
2.6 Conclusions ................................................................ 43
3 Estimation of the Transition Probabilities 49

3.1 Introduction ............................................. 49
3.2 Problem Formulation .................................... 50
  3.2.1 HMM State Space Model ............................ 50
  3.2.2 Conditional State Estimates and Information State Model .... 51
3.3 Estimation of Transition Probabilities .................... 52
  3.3.1 Least Squares ...................................... 53
  3.3.2 Extended Least Squares ............................ 55
  3.3.3 Recursive State Prediction Error Method .......... 56
3.4 Simultaneous Estimation ................................ 60
  3.4.1 Dual Cost Function Approach ........................ 61
3.5 Simulations ............................................. 64
  3.5.1 Implementation Considerations ....................... 64
  3.5.2 Simulation Results .................................. 64
3.6 Conclusions ............................................. 65

4 Almost Sure Estimation of Hidden Markov Models 71

4.1 Introduction ............................................. 71
4.2 Dynamics, Measure Change, and Martingale Properties .... 73
  4.2.1 Dynamics .......................................... 73
  4.2.2 Measure Change .................................... 74
  4.2.3 Transitions and Occupation Times .................... 75
  4.2.4 Strong Convergence ................................ 77
4.3 Conditional Mean Filters and Smoothers ................... 81
  4.3.1 On-Line Algorithms ................................ 81
  4.3.2 Algorithms Based on Smoothing ....................... 85
4.4 Estimation using Conditional Mean Estimates ............... 86
  4.4.1 Towards Global Convergence ......................... 87
  4.4.2 Global Convergence ................................ 88
  4.4.3 Reduced Complexity Algorithms ..................... 93
4.5 Simulations Studies ..................................... 94

x
4.5.1 Conditional Mean Estimates ................................. 94
4.5.2 Suboptimal Adaptive Schemes .............................. 95
4.5.3 Other Markov Chains ........................................ 96
4.5.4 A Comparison with Other Methods ......................... 96
4.5.5 Reduced Complexity Estimation ............................. 96
4.5.6 Threshold ..................................................... 97
4.5.7 More Complicated Approximations ......................... 97
4.5.8 High Noise and High Order Models ......................... 97
4.6 Conclusions ..................................................... 97

5 Estimation of Linear and Hybrid Linear Systems 103
5.1 Introduction ..................................................... 103
5.2 Dynamics ....................................................... 104
  5.2.1 Measure Change ........................................... 105
  5.2.2 Parameter Estimation - Full Observations ................. 106
  5.2.3 Almost Sure Convergence ................................ 107
5.3 Conditional Mean Estimates .................................. 109
  5.3.1 Preliminary Convergence Result ......................... 110
  5.3.2 Global Convergence Result ............................... 111
5.4 Hybrid Systems ................................................. 115
  5.4.1 Measure Change ........................................... 116
  5.4.2 Preliminary Results for Parameter Estimation ............ 117
  5.4.3 Almost Sure Convergence ................................ 118
  5.4.4 Conditional Mean Estimates ............................. 120
5.5 Simulations ..................................................... 122
5.6 Conclusions ..................................................... 123

6 Optimal HMM Filtering and Differential Decoding 125
6.1 Introduction ..................................................... 125
6.2 State Space Formulation ..................................... 127
  6.2.1 Signal Model ............................................. 127
  6.2.2 Channel Model ........................................... 128
6.2.3 State Space Signal Model ........................................ 128
6.3 Optimal HMM filter ..................................................... 131
6.3.1 State Space Representation ........................................ 131
6.3.2 Conditional Filter Formulation .................................... 133
6.4 Differential Receiver ................................................... 135
6.5 Decision Feedback Channel Estimation ............................ 136
6.5.1 Conditional HMM filter ............................................. 136
6.5.2 Conditional KF Channel Estimate ................................. 137
6.5.3 The Complete Algorithm ........................................... 139
6.6 Simulations ............................................................... 139
6.6.1 Differential Decoding With and Without Decision Feedback 139
6.7 Conclusion ............................................................... 140

7 Concluding Remarks and Future Work ............................... 143
7.1 Introduction ............................................................. 143
7.2 Summary of Contributions ............................................ 143
7.3 Future Research Directions ........................................... 147
# List of Figures

2.1 Empirical comparison of proposed schemes under assumption (3.23). ........ 43
2.2 Convergence to idealized state estimates. .................................. 44
2.3 Empirical comparison of proposed schemes. ............................... 44
2.4 Empirical comparison of the ELS scheme (3.32) and the RPE scheme of [6]. 45
2.5 Effect of stochastic approximation on convergence. ...................... 45
2.6 Parameter convergence of a six-state chain example. .................... 46
2.7 Parameter convergence in very high noise using Polyak acceleration. Averaging
   is performed over 1000 points ........................................... 46
2.8 Parameter convergence in very high noise using forgetting factors. ........ 47
2.9 The effect of variance correction factors. .................................. 47

3.1 Comparison of convergence rates ........................................... 67
3.2 Convergence in low noise .................................................. 67
3.3 Convergence of a fast chain ............................................... 68
3.4 Convergence of higher order chain ...................................... 68
3.5 Simultaneous estimation: A estimates .................................... 69
3.6 Simultaneous estimation: C estimates .................................... 69
3.7 Simultaneous estimation: Estimation error ............................... 70

4.1 Estimation error in $\hat{a}_{11}$ in medium noise .......................... 98
4.2 Estimation error in $\hat{C}_1$ in medium noise ............................ 98
4.3 Suboptimal Estimation of $\hat{C}_1$ in medium noise ...................... 99
4.4 Suboptimal Estimation of $\hat{C}_1$ in medium noise ...................... 99
4.5 Comparison of optimal and suboptimal filters ........................... 100
4.6 Estimation for unusual HMM .............................................. 100
4.7 The bias after 10000 points. ............................... 101
4.8 An empirical comparison ............................... 101

5.1 Estimation of $a_{21}$ ...................................... 124
5.2 Estimation of $a_{12}$ in higher noise ..................... 124

6.1 Block diagram of M-ary differential coherent receiver .... 135
6.2 Block diagram of differential receiver using channel estimates .... 136
6.3 Block diagram of coupled filter structure ................... 139
6.4 Improvement in BER performance due to decision feedback. ...... 141
Symbols and Abbreviations

Mathematical Symbols

- $\alpha_{k|\theta}, \alpha_{k|\theta-1,\theta}$ Unnormalized conditional estimates.
- $\theta$ Unknown parameters.
- $\gamma_k$ Step sequence or measure change.
- $\Gamma_k$ Measure change.
- $\kappa_k, \psi_k, \phi_k$ Gradients.
- $\lambda$ Model or random variable.
- $\lambda(\theta)$ Denotes parametrized model.
- $\lambda_k, \Lambda_k$ Measure change.
- $\pi_0$ Initial state value.
- $\Pi_k, \pi_k$ Measure change.
- $\phi$ Probability density function.
- $\psi$ Probability density function.
- $\mathbf{1}$ Column vector of all ones.
- $A, A, A(e_i)$ Transition matrix or stochastic transition matrix.
- $B, B(e_i)$ Noise mapping matrix.
- $B(y_k), B(y_k, \theta), b_k(i)$ Matrix of observation probabilities, elements.
- $C, C(e_i)$ Output mapping matrix or output matrix.
- $\mathbb{C}$ Complex Numbers.
- $D, D, D(e_i)$ Noise variance matrix.
- $e_i$ Unit vector.
- $\mathcal{F}_k$ A filtration.
- $g(.)$ A test function.
- $\mathcal{G}_k$ $\sigma$-field or filtration.
\(J_k, \mathcal{J}_k\)  
State transition count matrix.

\(M\)  
The model.

\(M_k, V_k\)  
Martingale increment.

\(N\)  
Number of states.

\(N_0\)  
Number of parameters.

\(O_k, \mathcal{O}_k\)  
State occupation count matrix.

\(N_k\)  
Normalization factor.

\(P_k\)  
An approximation of the Hessian.

\(R\)  
Real numbers.

\(T\)  
Number of time points.

\(T_k, \mathcal{T}_k\)  
State observation count matrix.

\(X_k\)  
Markov state at time \(k\).

\(\mathcal{X}\)  
Diagonal matrix with \(X\) on the diagonal.

\(V_k(\theta), V_k(\theta)^1, V_k(\theta)^2\)  
Cost functions.

\(i, j\)  
Index.

\(k, \ell\)  
Time index.

\(w_k, n_k\)  
Noise terms.

\(W\)  
Lyapunov function.

\(x_k\)  
Linear or hybrid linear state at time \(k\).

\(y_k\)  
Observations.

\(Y_k,\)  
A collection of observations up until time \(k\).

\(\mathcal{Y}_k\)  
\(\sigma\)-field or filtration of \(Y_k\).

\(z_k\)  
Observations.

\(\mathbb{Z}^+\)  
Positive integers.

Superscript and subscripts have the following possible meanings when applied to \(\eta\)

\(\{\eta\}\)  
Sequence or collection of \(\eta\).

\(\eta^*, \bar{\eta}\)  
Some other \(\eta\).

\(\eta^i, \eta^{(i)}\)  
The \(i\)th element of \(\eta\).

\(\eta_i, \eta_k\)  
The \(i\)th element of \(\eta\) or \(\eta\) at time \(k\).

\(\eta(i)\)  
The \(i\)th element of \(\eta\) or \(\eta\) evaluated at \(i\).

\(\eta^{ij}, \eta^{(ij)}, \eta_{ij}\)  
The \(ij\)th element of \(\eta\).
\( \hat{\eta}_k, \tilde{\eta}_k \)  
Estimate (or Conditional mean estimate) of \( \eta \) at time \( k \).

\( \hat{\eta}_{k|\ell} \)  
Estimate of \( \eta \) at time \( k \) given observations up until time \( \ell \).

\( \hat{\eta}_{k|M} \)  
Estimate of \( \eta_k \) given \( Y_\ell \) and the model \( M \).

\( \hat{\eta}_{k|\ell,\mathcal{M}} \)  
Estimate of \( \eta_k \) given \( Y_\ell \) and the sequence of models \( \mathcal{M}_k \).

\( \tilde{\eta}_k \)  
The error \( \tilde{\eta}_k = \eta - \eta \).

**Mathematical operations**

\( \langle \cdot, \cdot \rangle \)  
Inner product.

\( \text{diag}(\eta), \eta_{\text{diag}} \)  
The diagonal matrix formed by \( \eta \) or the diagonal of \( \eta \).

\( E[\cdot], E[\cdot|\cdot] \)  
Expectation operation.

\( := \)  
Defined as.

\( \langle \cdot \rangle_{2\pi} \)  
Modulo \( 2\pi \).

**Abbreviations**

DFE  
Decision Feedback Equalization.

DPMS  
Differential Phase Modulation Signal.

DPSK  
Differential Phase Shift Keying.

ELS  
Extended Least Squares.

EM  
Expectation Maximization.

HMM  
Hidden Markov Model.

IEEE  
Institute of Electrical and Electronic Engineers.

\( i.i.d. \)  
Individually, identically distributed.

ISI  
Inter-Symbol Interference.

KF  
Kalman Filter.

MAP  
Maximum a posteriori.

ODE  
Ordinary Differential Equation.

PAM  
Pulse Amplitude Modulation.

PSK  
Phase Shift Keying.

QAM  
Quadrature Amplitude Modulation.

RKL  
Recursive Kullback-Leibler.

RLS  
Recursive Least Squares.

RPE  
Recursive Prediction Error.

WGN  
White Gaussian Noise.
Chapter 1

Introduction

*Given me a thrill, says the reader*
*Give me a kick;*
*I don’t care how you succeed, or*
*What subject you pick.*

From *Fiction and the Reading Public* by Philip Larkin.

Daily we observe the world around us. The instruments we use to observe the world are imperfect. The world we observe is complicated. We do not expect these observations to perfectly reflect our world, neither from a philosophical nor a scientific viewpoint. Understanding the world we observe occupies much of our scientific effort.

The scientific terms we use to describe the behaviour we observe in our world include *signals* and *systems*. A system is an object, possibly unknown, in which interactions occur and it has observable behaviour (or signals). In this framework, input signals from the environment of the system are seen to cause the system to produce responses referred to as output signals.

From a scientific viewpoint, a system can be understood in terms of mathematical descriptions called models. These models provide a framework to aid understanding of the interactions between the inputs, the outputs and the system. There is a subtle, but crucial distinction between the concept of a model and the system. A model is no more than a practical tool used by scientists to attempt to describe a real world behaviour in mathematics.

Often the first aim of *system analysis* is to determine the mechanism that relates the input signals to the output signals and often mathematical models of this mechanism are
introduced.

Systematic tools have been developed for proposing and verifying models for systems. Even when system models can be developed from first principles using our theoretical understanding of the physical processes involved, these models need to be verified against real data obtained from experiments performed on the system. Often, a first principles theory is inadequate and models need to be developed and verified directly from experimental data. This type of system modelling is termed black-box modelling or system identification.

Once an accurate model has been obtained, various additional objectives can be satisfied, such as the closely related problems of signal processing and control. The task of signal processing is to infer from an observed output signal an estimate of the internal model quantities. Whereas, the task of control is to manipulate the system into producing a desired output sequence.

Implicit in the task of modelling an observed system is the belief that the system has some unobserved internal dynamics. For this reason, we allow ourselves to propose unknown internal structures and internal quantities to help explain the observations made. These internal variables are termed states. We use the assumed model and these states to help explain the relationship between the system’s input signals and the system’s output signals. Measurements are essential in refining and verifying our model for this relationship.

Often, there are many choices of internal variables and model structure, some overly simple, some overly complex. It is important to choose internal variables and model structures with enough complexity to reflect the behaviour of the system, but no more. This is reflected in Occam’s Razor: “The simplest explanation is best”.

1.1 Background and Motivation

Some of the many dynamics we observe in our world are discrete by their very nature. For example:

1. the number of sunny days in some city in a year;

2. whether a switch or machine is on or off;

3. base sequences in DNA;
4. bit sequences on computers or in telecommunications.

Some of these examples are intrinsically discrete-valued while others have continuous dynamics (such as the weather) that have been naturally categorized into discrete groups. It makes sense to develop models tailored to explain these types of dynamics. In this thesis we consider one type of discrete dynamics known as a hidden Markov model.

1.1.1 Hidden Markov Models

The basic concept of Markov chains was introduced by A.A. Markov in 1907 [58]. A Markov chain is a sequence of discrete valued states, in which the probability of a particular state occurring at a particular time depends only on the value of the state at the previous time. Interest in Markov processes was initially focused on their statistical properties and existence questions. Widespread understanding and application of the theory of Markov processes did not occur until much later with the advent of digital computers.

The term hidden Markov model (HMM) originated in the speech signal processing community (apparently due to L.P. Neuwirth [66]). HMMs have been extensively used in speech recognition tasks but the finite-discrete nature of HMM has resulted in them also becoming well known in the telecommunications area. Hidden Markov models in discrete time can be viewed as having an underlying unobservable state process. This state process is a Markov chain. The measurable outputs of an HMM are a probabilistic mapping from the present state of the Markov chain. The nature of this probabilistic mapping varies from application to application.

Hidden Markov models have been used in a variety of engineering and scientific applications including: speech processing and speech recognition systems [68], genetics and biochemistry, modelling time series sequences such as epileptic seizure counts [56], and ion channel currents in cell membranes [10]. There is substantial engineering literature on theory and methodology relating to Markov chains and HMMs: see Kemeny [35]; Seneta [69]; Elliott, Aggoun and Moore [18]; Le Gland [45]; Meyn and Tweedie [60]; and many articles in journals such as the IEEE Transactions on Signal Processing, IEEE Transactions on Control and Signal Processing.

This thesis concentrates on parameter estimation (or systems identification) for HMMs with a probabilistic mapping to a scalar continuous observation space. This is appropriate
for applications including: telecommunications [14, 15, 11, 79, 26], frequency tracking [76], ion currents measurements [10], as well as many others.

1.1.2 Linear Systems and Hybrid Linear Systems

Towards the end of this thesis, the results obtained for identification of HMMs are extended to identification of linear systems and hybrid linear systems.

In linear systems, unlike HMMs, there is an underlying continuous-valued process and an continuous-valued output process. Linear systems are the most commonly used system models and hence they are extensively treated in many texts. Rather than provide here an outline of the extensive treatment of linear systems the reader is referred to engineering literature on the topic, including: Anderson and Moore [1]; Kumar and Varaiya [41]; Ljung [51, 52]; as well as [7, 72, 73, 31]; and many articles in journals such as the IEEE Transactions on Signal Processing, IEEE Transactions on Automatic Control and Automatica.

Hybrid linear systems are an extension of linear systems. In this thesis, hybrid linear models are systems whose dynamics switch (or jump), according to a Markov chain, between various linear systems dynamics. Hybrid linear systems are a natural description for many physical processes. This includes systems that have a natural decomposition into discrete and continuous states as well as non-linear systems that are approximately linear in various regions. The reader is referred to articles in IEEE Transactions on Signal Processing and IEEE Transactions on Automatic Control, in particular [77, 37, 19].

1.1.3 System Identification

The parameter estimation task is the main subject considered throughout most of this thesis. Given a sequence of observations, we wish to estimate the unknown parameters of either a hidden Markov model, a linear system or a hybrid linear system. This type of problem is known as system identification. Identifying a system by observing its behaviour raises deep philosophical questions. What is the meaning of the estimated system? Is it the real system? Does it matter if it is not the real system?

In my opinion, before proceeding to perform system identification, the scientist or engineer must ask why he or she is modelling the system. Their answer to this question will directly influence how the subsequent model is interpreted.
There are four main reasons why models are identified.

1. **Modelling for control.** Here models are identified for the purpose of designing a controller, which when applied to the real plant will result in the controlled plant satisfying particular control objectives. The primary concern is that the control objectives are satisfied. The model needs only be good in the sense that it reflects the real system dynamics in the frequency ranges that are important for the control. Typically, a model must be good in particular frequency ranges and its short time prediction abilities are most important.

2. **Modelling for simulation.** Here models are identified so that the model can be used to produce long simulations of a real system that is inconvenient to measure. For example, behaviour that occurs over a long time scale (such as years) in nature can be studied in minutes via a simulation on a computer. Alternatively, expensive equipment can be protected from damage by simulation and experimentation on a computer rather than on the real system.

3. **Modelling for information (or hypothesis testing).** Here models are identified for the purpose of obtaining information directly from the system model. System models can be developed from first principles, conversely theory can be developed from estimated system models. Hypotheses can be tested against experimental data. The field of statistics is primarily concerned with this type of modelling. For example, in [56], Caesarean birth rates in a particular hospital were analysed and a model was identified which was used to imply that the proportion of Caesarean births to non-Caesarean births increased over time. Similarly, the system identification in [10] is motivated by a desire to understand the mechanisms of the human brain.

4. **Modelling for classification.** Here models are developed to allow classification of data sets into groups. First, a model is identified on a group of data sets for which the correct classification is known and then the model is used to classify unknown data sets. For example, given the billing history of a phone company client will the client switch to another carrier next month? Or, is the client committing phone fraud?

The *maximum likelihood* principle is a powerful statistical tool used to perform system
identification. The maximum likelihood estimate of a model arises at the maximum of the likelihood surface. The maximum likelihood criteria is philosophically satisfying because it estimates the model that is most likely to have produced the observations. Maximum likelihood estimates also have strong statistical properties.

Maximum likelihood estimation is typically very difficult to perform and hence other criteria are sometimes introduced, such as the prediction error cost and the Kullback-Leibler distance.

For linear systems, these criteria are often equivalent. However, for non-linear systems these criteria are generally not equivalent and it is not always clear which criterion is the most appropriate.

One well developed, general purpose technique for maximum likelihood estimation is the Expectation-Maximization (EM) algorithm [55]. The EM algorithm is an off-line locally convergent algorithm that locates maximum likelihood parameter estimates of a model on a data set after multiple passes through the data. The EM algorithm is attractive to users because it is intuitive and easy to formulate for new problems.

The main disadvantage of the EM algorithm is that it cannot be applied to on-line identification problems. These are problems in which new observations become available over time. It is desirable to include this information as it arrives to improve the ‘quality’ of the estimated model. Although off-line identification is important on its own, there are many situations were on-line identification is preferred or even required.

Generally, on-line algorithms have the advantage of requiring less memory and less overall computational effort at the cost of some optimality. Optimality is lost because the nature of the on-line problem requires that algorithms must forget past information as they process new information.

One well known approach for developing on-line algorithms is the recursive prediction error approach of Ljung [50, 51]. Recursive prediction error algorithms are developed to locate the local minimum of the prediction error cost function. This cost function measures the size of the error in the one-step-ahead predictions generated by candidate models. The recursive prediction error approach extends the concept of least squares from linear functions to non-linear functions. Because the cost function measures the error in predictions, it appears a natural criterion for the modelling for control problem.
1.1.4 A Brief History of Hidden Markov Model Estimation

A significant advance in theory for HMMs was the proposal of the optimal filter for continuous time HMMs by Wonham in 1965 [80]. This and the work by Elliott, Aggoun and Moore [18] provides solutions to the important state filtering problem. In the late 1960’s and early 1970’s, Baum and his colleagues, working on speech recognition problems, were interested in estimating the parameters of a discrete-time discrete-state HMM. They developed what later was shown to be an example of the EM algorithm [68]. The Baum-Welch algorithm is an off-line locally convergent scheme for obtaining maximum likelihood estimates of the HMM parameters.

The Baum-Welch algorithm has since facilitated work on problems as diverse as the estimation of models for cell membrane switching [10], frequency tracking [76] as well as Baum’s original problem of speech recognition [68].

A major limitation of this off-line multi-pass estimation algorithm is the ‘curse of dimensionality’ where the computational effort and memory requirements are proportional to the square of the number of Markov states and proportional to the length of the signal to be processed. In high noise, the number of passes required for convergence of parameter estimates can be high. For example, in some of our studies 1000 passes were required, see [10]. One avenue which has been investigated to reduce computational effort and memory requirements is through the use of on-line schemes.

Research efforts involving on-line schemes began in 1991 when Krishnamurthy and Moore proposed an on-line approximation of the EM algorithm [36] for the case when observations come from a discrete set. In 1992, Slingsby extended this algorithm for the case of continuous observations [71]. Then in 1993, Krishnamurthy and Moore proposed the Recursive Kullback-Leibler scheme (RKL) [39]. Each iteration of the parameter update equation in the RKL scheme has computational complexity of only $O(N_\theta)$ where $N_\theta$ is the number of parameters to be estimated. The disadvantage of the RKL scheme is that it converges noticeably slower than quadratic convergence; that is, the convergence rate is slower than $O(1/\sqrt{k})$, where $k$ is the number of data points.

In 1994, a scheme based on a recursive prediction error (RPE) approach was proposed by Collings, Krishnamurthy and Moore [13]. This RPE scheme provides improved convergence
rates. RPE schemes are known to be asymptotically efficient and provide quadratic convergence (of $O(1/\sqrt{k})$, where $k$ is the number of data points). However, this RPE scheme suffers the disadvantage of having, at each iteration of the parameter update equation, computational complexity of $O(N^2_\theta)$, which can be prohibitive for large $N_\theta$. This RPE scheme also suffers from numerical problems in low noise environments [16].

These RKL and RPE schemes have concentrated on identification of HMMs with continuous observations. Motivated to extend these results to the discrete observations case, Le Gland and Mével proposed algorithms based on minimizing the log-likelihood function [45]. In 1996, parameter estimation via time-discretization of continuous-time filters were proposed by James, Krishnamurthy and Le Gland [43]

### 1.2 Outline of Thesis Contributions

The technical results in this report are presented in roughly the chronological order in which they were obtained. Table 1 outlines the contributions made in this thesis in the area of parameter estimation for HMMs and linear systems. The table highlights the way in which the techniques presented in this thesis supersede existing techniques. The arrows in the table highlight the evolution of the techniques and their relationships to one another.

<table>
<thead>
<tr>
<th></th>
<th>Off-line Local</th>
<th>On-line local</th>
<th>On-line Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>Healy-Westmacott EM</td>
<td>RPE</td>
<td>[Ch. 5]</td>
</tr>
<tr>
<td>HMM</td>
<td>EM</td>
<td>Variety of solutions and [Ch. 2 &amp; 3]</td>
<td>[Ch. 4]</td>
</tr>
</tbody>
</table>

A more detailed outline of the results presented in this thesis is given below.
Prediction error algorithm for estimation of Hidden Markov Models parameters

On-line recursions for estimating HMM model parameters are important in many applications. In Chapters 2 and 3, algorithms for estimating the parameters of an HMM using a recursive prediction error approach are derived. The presented algorithms are computationally attractive while exhibiting quadratic convergence rates. Convergence results and convergence rate results are established using an ODE approach.

Suboptimal extended least squares algorithms are also presented in these chapters and convergence results are established in idealized situations.

The algorithm proposed by Collings, Krishnamurthy and Moore, in [13], suffers from convergence difficulties in low noise environments. An explanation of why the typical least squares cost, used in [13], is inappropriate for estimation of the transition probability matrix is given in Chapter 3.

The algorithms in Chapters 2 and 3 are developed by proposing new cost functions that alleviate this difficulty in low noise as well as exploiting the discrete-valued nature of HMMs.

While the algorithms of Chapters 2 and 3 offer improvements over the existing techniques, these algorithms are only locally convergent and in some situations do not converge from poor initializations.

Almost Sure Parameter Estimation for Hidden Markov Models

Globally convergent estimation schemes are desirable, since they converge from any initial parameter estimate. This motivates our work in Chapter 4 where almost surely convergent algorithms for estimating HMM parameters are presented. Convergence results are established using martingale convergence results, the Kronecker lemma, and an ODE approach. Suboptimal algorithms are presented but convergence results are not established.

The algorithms presented in Chapter 4 look similar to the well known locally convergent Baum-Welch EM algorithms and are linked to the RPE algorithms of Chapters 2 and 3 through a Lyapunov function.

Almost Sure Parameter Estimation for Linear and Hybrid Linear Systems

In Chapter 5, almost surely convergent algorithms for estimating parameters of linear systems and hybrid linear system are presented. Partial convergence results are established for
both estimators using martingale convergence results and the Kronecker lemma. Complete convergence results for the parameter estimator for linear systems are established using an ODE approach.

**Differentially Encoded Signals**

Message sequences in digital communications are generally assumed to be individually and identically distributed. When convolutionally coded, the coded sequences are Markov chains. This motivates the work in Chapter 6 where new techniques for demodulating differentially encoded signals are proposed.

Here the message sequence is assumed to be a Markov chain. Hence, the problem of demodulating a differentially encoded signal is considered as an application of our HMM filtering ideas. A signal model of the modulation system and communication channel is proposed and optimal HMM filters are presented. The use of decision feedback is also investigated to aid demodulation performance. Simulation examples are presented which demonstrate that decision feedback improves receiver performance. The performance gains are appealing but more work needs to be done before this approach can be considered as a practical alternative.

### 1.3 Review of Various Mathematical Tools

Some mathematical preliminaries are presented in this section. Some of the tools presented here will be repeated in detail when used in this thesis. Others are provided here to familiarize the reader with known theory that is referred to but not explicitly quoted in this thesis.

The reader is referred to Billingsley [9] for an introduction into probability and measure theory; to Doob [17] for an introduction into stochastic processes; to Elliott, Aggoun and Moore [18] for an introduction into hidden Markov models; and to [51, 52, 72] for general introductions into system identification.

**Hidden Markov Model**

The HMM notation presented below is used throughout most of this thesis. There are slight deviations in some chapters but these deviations will be clear when they occur. Let $X_k$ be a discrete-time homogeneous, first order Markov process, belonging to a finite set. The state space, $X$, *without loss of generality*, can be identified with a set of unit vectors, $S =$
\{e_1, e_2, ..., e_N\}, \ e_i = (0, ..., 0, 1, 0, ..., 0)' \in \mathbb{R}^N \text{ with } 1 \text{ in the } i\text{th position}. \text{ The transition probability matrix is}

\[ A = (a_{ij}) \ 1 \leq i, j \leq N \text{ where } a_{ij} = P(X_{k+1} = e_i | X_k = e_j) \]  \quad (1.3.1)

so that

\[ E[X_{k+1} | X_k] = AX_k \] \quad (1.3.2)

where \( E[] \) denotes the expectation operator. We also denote \( \{\mathcal{F}_\ell, \ell \in \mathbb{Z}^+\} \) the complete filtration generated by \( X \); That is, for any \( k \in \mathbb{Z}^+ \), \( \mathcal{F}_k \) is the complete filtration generated by \( X_\ell, \ell \leq k \). For a brief introduction of the concept of filtration see [18, Appendix A].

The dynamics of \( X_k \) are given by the state equation

\[ X_{k+1} = AX_k + M_{k+1} \] \quad (1.3.3)

where \( M_{k+1} \) is a \( (A, \mathcal{F}_k) \) martingale increment, in that \( E[M_{k+1} | \mathcal{F}_k] = 0 \).

We assume \( X_k \) is hidden; that is, \( X_k \) is indirectly observed by measurements \( y_k \) in a continuous range, \( R \). The observation process \( y_k \) is assumed to be scalar (for simplicity of presentation only) and to have the form

\[ y_k = C'X_k + w_k = X_k'C + w_k \in R \] \quad (1.3.4)

where \( C \in \mathbb{R}^N \) is the vector of state values of the Markov chain, and \( w_k \) is independent and identically distributed (i.i.d.), with zero mean and Gaussian density, i.e., \( w_k \sim N(0, \sigma_w^2) \) and \( E[w_{k+1} | \mathcal{F}_k \vee \mathcal{Y}_k] = 0 \) where \( \mathcal{Y}_i \) is the complete filtration generated by \( y_k, k \leq l \). We also define \( Y_k := (y_0, ..., y_k) \).

We shall define the vector of parametrized probability densities (or symbol probabilities) as \( b_k = (b_k(i)), \text{ for } b_k(i) := P[y_k | X_k = e_i] \). In the special case when \( w_k \) is \( N(0, \sigma_w^2) \) we can write,

\[ b_k(i) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp \left( \frac{-(y_k - C'e_i)^2}{2\sigma_w^2} \right). \] \quad (1.3.5)

The initial state probability vector for the Markov chain is denoted \( \pi = (\pi_i) \) where \( \pi_i = P(X_1 = e_i) \). The HMM is denoted as \( \lambda = (A, C, \pi, \sigma_w^2) \).

**Linear Models**

The notation presented below is used in this thesis in the discussion involving partially observed linear stochastic systems.
Consider a probability space \((\Sigma, \mathcal{F}, P)\); suppose \(\{x_t\}, t \in \mathbb{Z}^+\) is a discrete-time linear stochastic process, taking values in \(\mathbb{R}^N\), with dynamics given by

\[
x_{k+1} = Ax_k + Bw_{k+1}, \quad x_0 \in \mathbb{R}^{N \times 1}.
\]

Here \(k \in \mathbb{Z}^+, A \in \mathbb{R}^{N \times N}, B \in \mathbb{R}^{N \times 1}\) and \(\{w_t\}, t \in \mathbb{Z}^+\), is a sequence of independent and identically distributed \(N(0,1)\) scalar random variables. The probability density of the noise, \(Bw_t\), is denoted as \(\psi\).

The state process \(x_t, t \in \mathbb{Z}^+\), is observed indirectly via the scalar observation process \(\{y_t\}, t \in \mathbb{Z}^+\), given by

\[
y_k = Cx_k + Du_k.
\]

where \(C \in \mathbb{R}^{1 \times N}, D \in \mathbb{R}\) for \(k \in \mathbb{Z}^+\) and \(\{v_t\}, t \in \mathbb{Z}^+\), is a sequence of independent and identically distributed \(N(0,1)\) scalar random variables. The probability density of the noise, \(Du_t\), is denoted as \(\phi\). We further assume that \(x_0, \{v_t\}\) and \(\{w_t\}\) are mutually independent. The linear system is denoted \(\lambda = \lambda(A, B, C, D, x_0)\).

**Prediction Error Algorithms**

The algorithms in Chapters 2 and 3 are developed by modifying the standard techniques presented in Ljung [50, 51] for estimating systems. Here is a brief outline of the prediction error approach to developing algorithms.

In the simplest case, consider a linear system with an assumed signal model as below (1.3.8), and parametrized by \(C\).

\[
y_k = CX_k + w_k
\]

where \(y_k \in \mathbb{R}, C \in \mathbb{R}^{1 \times N}, X_k \in \mathbb{R}^{N \times 1}\) and \(w_k \sim N(0, \sigma_w^2)\). The value of \(C\) can be estimated from measurements \(\{y_k\}\) and \(\{X_k\}\) using the least squares (LS) algorithm.

\[
\begin{align*}
\hat{C}_k &= \hat{C}_{k-1} + P_k X_k (y_k - \hat{C}_{k-1} X_k) \\
P_k^{-1} &= P_{k-1}^{-1} + X_k X_k' 
\end{align*}
\]

where \(\hat{C}_k\) denotes the estimate of \(C\) at time \(k\). This problem is made more difficult if \(X_k\) is not measured directly but is only available via estimates \(\hat{X}_k\).

A naive approach commonly used, termed extended least squares (ELS), is to simply replace the state in (1.3.9) by a state estimate. No convergence results are available for the ELS algorithm.
In [51, 50] Ljung presents algorithms for which local convergence can be guaranteed, see also the work of Moore and Weiss[63]. Ljung’s work is based on minimisation of the prediction error cost function. The prediction error cost can be interpreted as an error induced by errors in the model. For this reason, minimisation of the prediction error cost function $V_k(\theta)$ is used as a criteria for estimating model parameters as follows:

$$\hat{\theta}_k = \arg\min_{\theta} \{ V(\theta) = E[y_k - \hat{y}_{k|k-1, \theta}]^2 \}$$

where $\hat{y}_{k|k-1, \theta} := E[y_k|y_0, \ldots, y_{k-1}, \theta]$. A common technique used to minimize this cost function is known as the Gauss-Newton gradient search approach. That is,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \gamma_k \left( \left. \frac{d^2 V(\theta)}{d\theta^2} \right|_{\theta = \hat{\theta}_k} \right)^{-1} \left. \frac{d V(\theta)}{d\theta} \right|_{\theta = \hat{\theta}_k}$$

After several assumptions and several steps, the following practical algorithm can be proposed, see [51] for more details.

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \gamma_{k+1} R_{k+1}^{-1} \left( \left. \frac{d \hat{y}_{k|k-1, \theta}^{-1} \hat{y}_{k-1, \theta}^{-1} y_k - \hat{y}_{k|k-1, \theta}^{-1}}{d\theta} \right|_{\theta = \hat{\theta}_k} \right) \hat{y}_{k|k-1, \theta}^{-1} \hat{y}_{k-1, \theta}^{-1} \hat{\theta}_k$$

(1.3.10)

where $\hat{y}_{k|k-1, \theta}^{-1} \hat{y}_{k-1, \theta}^{-1} y_k - \hat{y}_{k|k-1, \theta}^{-1} \hat{y}_{k-1, \theta}^{-1} \hat{\theta}_k$. Note that (1.3.10) is equivalent to (1.3.9) for linear systems.

The techniques of Ljung were developed for linear and non-linear systems with variables in the continuous range. The discrete-state nature of Markov models means that direct application of Ljung’s techniques is not appropriate without reformulation of the task. In this thesis the standard technique is modified by considering new cost functions. For example, in Chapter 3, instead of considering an observation prediction error cost, the following state prediction error cost is considered:

$$\tilde{V}_k(\theta) = E[X_k - \tilde{X}_{k|k-1, \theta}]^2$$

(1.3.11)

where $\tilde{X}_{k|k-1, \theta} := E[X_k|y_0, \ldots, y_{k-1}, \theta]$. Recursions for minimisation of this cost function are developed in parallel manner to Ljung’s approach.

**Expectation-Maximization Algorithm**

The EM has been successfully applied to a number of off-line estimation problems. Here, is a preliminary introduction of the EM algorithm.
Let \{y\} denote some observed data set having a probability density function postulated as \( p_y(y|\theta) \) where \( \theta \) is a vector of unknown parameters in some space \( \Omega \). The EM algorithm is an iterative procedure for locating maximum likelihood estimates in situations where, given some extra unknown data, maximum likelihood estimation is relatively straightforward. Let \{x\} denote this missing data. The problem of estimating \( \theta \) from \{y\} is termed the incomplete data problem while the problem of estimating \( \theta \) from \{y\} and \{x\} is termed the complete data problem.

Let \( p_x(x|\theta) \) denote the probability density function of the unknown random variable \( x \). Then, the log-likelihood function for the complete data problem is given by

\[
\log L_c(\theta) = \log p_x(x|\theta)
\]

(1.3.12)

The EM algorithm approaches the incomplete data problem indirectly by proceeding to solve the complete data problem. This is appropriate if the complete data problem is relatively easy. The complete data log-likelihood function \( \log L_c(\theta) \) is unobservable so the conditional expectation of \( \log L_c(\theta) \) with respect to \{y\} and \( \hat{\theta}_{k-1} \) is used instead of \( \log L_c(\theta) \), where \( \hat{\theta}_{k-1} \) is the latest estimate of \( \theta \).

The EM algorithm is a two-step iterative procedure where the next estimate \( \hat{\theta}_k \) is calculated from the previous estimate \( \hat{\theta}_{k-1} \) as follows

**E step** : Calculate

\[
Q(\theta, \hat{\theta}_{k-1}) := E[\log L_c(\theta)|\{y\}, \hat{\theta}_{k-1}]
\]

**M step** : Choose \( \hat{\theta}_k \) such that

\[
\hat{\theta}_k = \arg \max_{\theta \in \Omega} Q(\theta, \hat{\theta}_{k-1})
\]

The E and M steps are repeated until convergence has occurred.

It has been shown that the value of the likelihood function at each step of the EM algorithm never decreases. Under mild conditions the EM algorithm converges to a local maximum of the likelihood surface [55].

The off-line algorithm given in [68] for estimating the parameters of an HMM is an example of an EM algorithm.

**E step** : Calculate

\[
E\left[ \sum_{k=1}^{T-1} X_k^i X_{k+1}^j |y_0, \ldots, y_T, \lambda \right],
\]
\[ E \left[ \sum_{k=1}^{T-1} X_k^i y_0, \ldots, y_T, \hat{\lambda}_\ell \right] \quad \text{and} \quad E \left[ \sum_{k=1}^{T-1} X_k^i y_k, y_0, \ldots, y_T, \hat{\lambda}_\ell \right] \]

where \( X^i_k \) is the \( i \)th element of \( X_k \) and \( \hat{\lambda}_\ell \) is the model estimate on the \( \ell \)th pass through the data.

**M step**: Calculate the parameter estimates \( \hat{\lambda}_{t+1} \) and \( \hat{C}_{t+1} \) at recursion step \( \ell + 1 \). Set,

\[
\hat{A}_{t+1}^{ij} = \frac{E[\sum_{k=1}^{T-1} X^i_k X^j_k | y_0, \ldots, y_T, \hat{\lambda}_\ell]}{E[\sum_{k=1}^{T-1} X^i_k | y_0, \ldots, y_T, \hat{\lambda}_\ell]} \quad \text{for} \quad 1 \leq i, j \leq N
\]

and

\[
\hat{C}_{t+1}^i = \frac{E[\sum_{k=1}^{T-1} X^i_k y_k | y_0, \ldots, y_T, \hat{\lambda}_\ell]}{E[\sum_{k=1}^{T-1} X^i_k | y_0, \ldots, y_T, \hat{\lambda}_\ell]} \quad \text{for} \quad 1 \leq i \leq N
\]

where \( \hat{A}_{t+1}^{ij} \) is the \( ij \)th element of \( \hat{A}_{t+1} \) and \( \hat{C}_{t+1}^i \) is the \( i \)th element of \( \hat{C}_{t+1} \). The model estimate \( \hat{\lambda}_{t+1} \) can be formed from \( \hat{A}_{t+1} \) and \( \hat{C}_{t+1} \).

Repeat E and M steps until convergence occurs.

**Healy-Westmacott Procedure**

An interesting identification problem which turns out to be related to the EM algorithm is the least squares with missing data problem. In this problem it is supposed that data is generated by a linear model

\[ y_j = \beta_0 + \beta x_j + e_j, \quad j = 1, \ldots, n, \]

where \( e_j \) is a Gaussian error and \( \beta \) are unknown model parameters. This would be a simple least squares problem, if not for the fact that some of the observations \( y_j \) are missing. An approach for this sort of problem was suggested by Healy and Westmacott \[33\] in 1959. It went as follows:

1. Estimate (or guess) values for the missing \( y_j \).
2. Solve least squares problem using real data + estimated data to produce a model estimate.
3. Use the estimate model to estimate the missing data.
4. Return to Step 2 until convergence occurs.
This is clearly a particular example of the EM algorithm, where Step 2 is the maximization step and Step 3 is the expectation step.

An alternative method for handling the least squares missing data problem is using the recursive prediction error approach of Ljung [51]. This RPE approach is taken in [13] to identify the parameters of an HMM. It is not clear if the authors of [13] were aware of the Healy-Westmacott procedure for handling missing data.

**Idempotent Property**

Throughout this thesis the idempotent property of Markov chains is exploited to reduce computational effort.

The **idempotent** property is

\[ X_jX_i^t \in S^* \]

where \( S^* = \{e_1e_1',\ldots,e_ie_j',\ldots,e_Ne_N'\} \), that is, the set of matrices which are zero everywhere except the \( ij \)th element which is 1. In particular, it follows that \( X_kX_k' = diag(X_k) \) where \( diag(X_k) \) is the diagonal matrix with \( X_k \) on its diagonal and hence non-linear functions of \( X_k \) are linear functions of \( X_k \).

**Convergence Results**

To aid the reader we introduce here the concepts of *almost sure* convergence and *mean square* convergence.

We say that in a probability measure space \( (\Omega, \mathcal{F}, P) \) a result holds almost surely (a.s) if it holds with probability 1, or equivalently that is holds for \( \omega \) in an \( \mathcal{F} \)-set of probability 1, where \( \omega \in \Omega \). We are concerned with convergence results so consider the following illustration. Let \( x, x_1, x_2, \ldots \) be random variables. If

\[
\lim_{k \to \infty} x_k(\omega) = x(\omega)
\]

for \( \omega \) in an \( \mathcal{F} \)-set of probability 1, then

\[
\lim_{k \to \infty} x_k = a \text{ almost surely or a.s.}
\]

A weaker convergence condition is that

\[
\lim_{k \to \infty} P[|x_k(\omega) - x(\omega)|^2 \geq \epsilon] = 0
\]
which is called convergence in mean square (and is a convergence in probability result). Note that almost sure convergence implies mean square convergence but mean square convergence does not imply almost sure convergence.

Two convergence results that are used often in this thesis are the martingale convergence theorem and the Kronecker lemma [64, 59, 53]. These two results are stated below.

**Theorem:** (Martingale Convergence)

Let \( \{M_k, F_k\} \) be a martingale, that is, \( F_k \) is an increasing sequence of \( \sigma \)-algebras, \( M_k \) is \( F_k \)-measurable, and \( E[M_k+1|F_k] = M_k \) a.s. for all \( k \). If \( \sup_k E|M_k|^p < \infty \) for some \( p \geq 1 \), then \( \{M_k\} \) converges to an a.s. finite random variable. (When this holds with \( p = 2 \), the martingale is said to be square integrable.)

**Theorem:** (Kronecker’s Lemma)

Let \( \{x_k\} \) and \( \{r_k\} \) be two real valued sequences satisfying

\[
    r_k > 0, \quad \lim_{k \to \infty} r_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{x_k}{r_k} < \infty.
\]

Then,

\[
    \lim_{N \to \infty} \frac{1}{r_N} \sum_{k=1}^{N} x_k = 0.
\]

### 1.4 Definitions

Definitions adopted by researchers are often not uniform, so key and controversial terms are defined in this thesis as follows:

- **Output Mapping**: The vector (or matrix) which maps the state to noise-less output.
- **Stochastic matrix**: Matrix of probabilities: Each element is nonnegative and column sums must equal 1.
- **State Values**: The values in the output mapping vector.
- **Filtration**: \( \sigma \)-field generated by a process, see [18].
- **Complete Filtration**: Filtration augmented with all subsets of events of probability zero.
- **Almost Sure**: The set on which the event does not occur has measure zero, see [9].
- **Hybrid Linear System**: A linear system whose parameters change according to a Markov chain.
1.5 Delimitation of Scope and Key Assumptions

The investigation of HMM parameter estimation presented in this thesis is primarily limited to presentation, verification and simulation of algorithms for estimating parameters of time-invariant, first-order Markov systems with additive white Gaussian noise.

The investigation of linear system parameter estimation is limited to presentation, verification and simulation of algorithms for estimating parameters of partially observed linear system and hybrid linear system in Chapter 5.

It is an implicit assumption that the true model is a member of the model set. However, this may be relaxed in some places.
Chapter 2

Estimation of the State Values

2.1 Introduction

Hidden Markov models (HMMs) are a powerful tool in the field of signal processing [35, 18] with application to speech processing [68], digital communication systems [14] and biological signal processing [10]. The major limitation of schemes for the estimation of HMM parameters revolve around computation and memory requirements.

Hidden Markov models in discrete time can be viewed as having a state $X_k$ at time $k$ belonging to a discrete set, without loss of generality denoted as $S = \{e_1, e_2, ..., e_N\}$, where $e_i$ is a vector that is zero everywhere excepting the $i$th element which is 1. There are transitions between states described by fixed probabilities which form a matrix $A = (a_{ij})$ where $a_{ij}$ is the probability of transferring from state $e_j$ to state $e_i$. The model measurements are an output mapping from the Markov states, $\theta X_k$ contaminated by additive noise. The corruption of the output mapping is the reason the model is termed hidden.

The Expectation-Maximization (EM) algorithm [68] is a popular off-line locally convergent scheme for obtaining maximum likelihood estimates of the HMM parameters. A major limitation of off-line multi-pass estimation schemes is the ‘curse of dimensionality’ where the computational effort and memory requirements are in proportion to the square of the number of Markov states and proportional to the length of the signal to be processed. One avenue to improve the computational and memory requirements would appear to be through the investigation of online schemes. It should also be said that in learning the model parameters in a multi-pass arrangement, convergence rates are linear, meaning of order $1/N$ with respect to the number of passes, $N$, through the data.
The two notable examples of online adaptive schemes for HMM parameters estimation are the Recursive Kullback-Leibler scheme (RKL) [39] and the recursive prediction error scheme (RPE) [13]. The RKL scheme converges linearly, each iteration of the parameter update equation has computational complexity of $O(N_\theta)$ where $N_\theta$ is the number of parameters to be estimated. The RPE scheme [13] was developed with the aim to provide improved convergence rates. This scheme is known to be asymptotically efficient and provide quadratic convergence (of $O(1/\sqrt{k})$) where $k$ is the number of points. However, each iteration of the parameter update equation has computational complexity of $O(N_\theta^2)$, which can be prohibitive for large $N_\theta$.

The key contributions of this chapter are the proposal of several new online schemes for HMM parameter estimation, based on extended least squares (ELS) and recursive prediction error (RPE) concepts with the ELS approach rationalized via martingale convergence results, and convergence results shown for the RPE schemes via an ordinary differential equation (ODE) approach. The best of these new schemes are based on a least squares prediction error index that uses \textit{a posteriori} estimates rather than prediction estimates.

A typical application in the simplest of contexts, under study in a later chapter, is the demodulation of coded QAM signals with known transition probabilities in a noisy fading channel. The state transition probabilities and channel noise statistics would be assumed known but the channel gain and phase changes unknown, and possibly time varying. The problem of estimating the transition probabilities is considered in the following chapter.

The chapter is organized as follows: In Section 2.2, we formulate the signal model and introduce an information state model. In Section 2.3, we introduce first the simplest case of the adaptive estimation task, namely when the state sequence is measured directly, and then apply the least squares approach familiar in linear system identification. When the state sequence is not measured directly, the least squares approach leads to the ELS algorithms. Some convergence results are presented. In Section 2.4, we generalize the ELS algorithms by introducing RPE recursion schemes, with new search directions and ODE convergence results are presented. A new cost function, which is suggest by the least squares approach, and the \textit{a posteriori} weighted RPE scheme is also presented. In Section 2.5, implementation considerations are discussed and simulation examples are given. Finally, some conclusions are presented in Section 2.6.
2.2 Problem Formulation

In this section we describe the HMM signal model in state space form, discuss its parametrization and reformulate it as an information state model.

2.2.1 HMM State Space Model

Let $X_k$ be a discrete-time homogeneous, first order Markov process, belonging to a finite set. The state space, $X$, without loss of generality, can be identified with a set of unit vectors, $S = \{e_1, e_2, \ldots, e_N\}$, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^t \in \mathbb{R}^N$ with 1 in the $i$th position. The transition probability matrix is

$$A = (a_{ij}) \quad \text{for} \quad 1 \leq i, j \leq N \quad \text{where} \quad a_{ij} = P(X_{k+1} = e_i | X_k = e_j) \quad (2.2.1)$$

so that

$$E[X_{k+1} | X_k] = AX_k \quad (2.2.2)$$

where $E[.]$ denotes the expectation operator. We also denote $\{\mathcal{F}_l, l \in \mathbb{Z}^+\}$ the complete filtration generated by $X$, that is, for any $k \in \mathbb{Z}^+$, $\mathcal{F}_k$ is the complete filtration generated by $X_l$, $l \leq k$. For a brief introduction of the concept of filtration see [18, Appendix A].

**Lemma 2.2.1** The dynamics of $X_k$ are given by the state equation

$$X_{k+1} = AX_k + M_{k+1} \quad (2.2.3)$$

where $M_{k+1}$ is a $(A, \mathcal{F}_k)$ martingale increment, in that $E[M_{k+1} | \mathcal{F}_k] = 0$.

**Proof** [18, 64]

$$E[M_{k+1} | \mathcal{F}_k] = E[X_{k+1} - AX_k | X_k, A] = E[X_{k+1} | X_k, A] - AX_k = 0.$$

We assume $X_k$ is hidden, that is, indirectly observed by measurements $y_k$ in a continuous range $R$. The observation process $y_k$ is assumed to be scalar (for simplicity of presentation only) and to have the form

$$y_k = \theta' X_k + w_k = X_k' \theta + w_k \in R \quad (2.2.4)$$
where $\theta \in \mathbb{R}^N$ is the vector of state values of the Markov chain. We also define $Y_k := (y_0, \ldots, y_k)$. We assume $w_k$ is i.i.d., with zero mean and Gaussian density, i.e. $w_k \sim N(0, \sigma_w^2)$ and $E[w_{k+1}|\mathcal{F}_k \lor \mathcal{Y}_k] = 0$ where $\mathcal{Y}_i$ is the complete filtration generated by $y_k$, $k \leq i$. We denote the complete filtration generated by $\{X_1, \ldots, X_k, y_1, \ldots, y_k\}$ as $\mathcal{G}_k$.

We shall define the vector of parametrized probability densities (or symbol probabilities) as $b_k = (b_k(i))$, for $b_k(i) := P(y_k|X_k = e_i)$. In the special case as here when $w_k$ is i.i.d. and $N(0, \sigma_w^2)$ we can write,

$$b_k(i) = \frac{1}{\sqrt{2\pi \sigma_w^2}} \exp \left( \frac{-(y_k - \theta' e_i)^2}{2\sigma_w^2} \right). \quad (2.2.5)$$

We can also write the initial state probability vector for the Markov chain $\pi = (\pi_i)$ is denoted by $\pi_i = P(X_1 = e_i)$. The HMM is denoted $\lambda = (A, \theta, \pi, \sigma_w^2)$.

Model Parametrization

For simplicity, in this chapter we shall be considering the problem of estimating unknown state values, assuming knowledge of $A$, $\pi$ and $\sigma_w^2$, as in communication channels with known coding. Let $\lambda$ be parametrized by an unknown vector $\theta$ so that the parametrized HMM is denoted by,

$$\lambda(\theta) = (A, \theta, \pi, \sigma_w^2).$$

2.2.2 Information State Model

Let $\hat{X}_{k|k,\theta}$ and $\hat{X}_{k|k-1,\theta}$ denote the conditional filtered state estimate and one-step-ahead state prediction of $X_k$ at time $k$, given measurements $Y_k$, up until time $k$ and the parameter $\theta$ i.e,

$$\hat{X}_{k|k,\theta} := E[X_k|\mathcal{Y}_k, \theta], \quad \hat{X}_{k|k-1,\theta} := E[X_k|\mathcal{Y}_{k-1}, \theta]. \quad (2.2.6)$$

Note that $\hat{X}_{k|k-1,\theta} = A\hat{X}_{k-1|k-1,\theta}$.

It is often convenient to work with an unnormalized conditional estimates (or the so-called “forward” variables), $\alpha_{k|k,\theta}$ and $\alpha_{k|k-1,\theta}$ which are defined as $\alpha_{k|k,\theta} = (\alpha_{k|k,\theta}(i))$, for $\alpha_{k|k,\theta}(i) := P(Y_k, X_k = e_i|\theta)$ and $\alpha_{k|k-1,\theta} = A\alpha_{k-1|k-1,\theta}$. These unnormalized conditional estimates are computed using the following “forward” recursion:

$$\alpha_{k+1|k+1,\theta} = B(y_{k+1}, \theta) A\alpha_{k|k,\theta}, \quad \alpha_{k+1|k,\theta} = AB(y_k, \theta) \alpha_{k|k-1,\theta} \quad (2.2.7)$$
where $B(y_k, \theta) = \text{diag}([b_1(1), \ldots, b_k(N)])$ and $\text{diag}(X)$ is the diagonal matrix with $X$ on its diagonal.

We can now write the conditional filter estimate and one-step-ahead prediction, from the unnormalized conditional estimates, as

$$\hat{X}_{k|k, \theta} = (\alpha_{k|k, \theta}^{\perp})^{-1} \alpha_{k|k, \theta}$$
$$\hat{X}_{k|k-1, \theta} = (\alpha_{k|k-1, \theta}^{\perp})^{-1} \alpha_{k|k-1}$$

(2.2.8)

where $(\cdot, \cdot)$ is an inner product and $\mathbf{1}$ is the column vector containing all ones.

**Parametrized Filtered Estimate**

We now seek to express the observations $y_k$ in terms of the conditional filter estimate at time $k$.

**Lemma 2.2.2 [13]** The conditional measurements $y_k$ are given by

$$y_k = \theta^T \hat{X}_{k|k, \theta} + n_k$$

where

$$n_k = \theta^T [X_k - \hat{X}_{k|k, \theta}] + w_k.$$  

The parametrized filtered estimates, $\hat{X}_{k|k, \theta}$, are given by

$$\hat{X}_{k|k, \theta} = N_k B(y_k, \theta) A \hat{X}_{k-1|k-1, \theta}$$
$$y_k = \theta^T \hat{X}_{k|k, \theta} + n_k$$

(2.2.9)

where $N_k = (B(y_k, \theta) A \hat{X}_{k-1|k-1, \theta} ^{\perp})^{-1}$ is a normalization factor.

**Parametrized One-Step-Ahead-Prediction**

We now seek to express the observations $y_k$ in terms of a prediction based on the conditional filter estimates at time $k - 1$.

**Lemma 2.2.3 [13]** In the above notation, the measurements $y_k$ are given by

$$y_k = \theta^T \hat{X}_{k|k-1, \theta} + n_k$$

where

$$n_k = \theta^T [X_k - \hat{X}_{k|k-1, \theta}] + w_k$$

(2.2.10)

and $n_k$ is a $(\theta, Y_{k-1})$ martingale increment and is white.
Proof [13] Following the standard arguments since \( \hat{X}_{k|k-1|\theta} \) is measurable with respect to \( \theta, Y_{k-1} \), \( E[w_{k+1}|Y_k] = 0 \) and \( E[M_{k+1}|Y_k] = 0 \), then
\[
E[n_k|\theta, Y_{k-1}] = \theta^T \hat{X}_{k|k-1,\theta} - \theta^T \hat{X}_{k|k-1,\theta} = 0.
\]
□

In summary, the parametrized predictor based signal model, for an HMM parameter \( \theta \) and with states \( \hat{X}_{k|k-1,\theta} \) is given by
\[
\hat{X}_{k+1|k,\theta} = N_k AB(y_k, \theta) \hat{X}_{k|k-1,\theta}
\]
\[
y_k = \theta^T \hat{X}_{k|k-1,\theta} + n_k
\]
(2.2.11)
where \( n_k \) is a \((\theta, Y_{k-1})\) martingale increment and \( N_k = (AB(y_k, \theta) \hat{X}_{k|k-1,\theta}, 1)^{-1} \) is a normalization factor.

We now proceed to consider the problem of estimating \( \theta \) given a sequence of observations, \( Y_k \).

2.3 Least Squares and Extended Least Squares

This section has two parts. In the first part, to introduce the problem we consider the simplified adaptive estimation task for the case when \( X_k \) is measured. The familiar least squares algorithm from linear system identification theory is presented and the idempotent nature of indicator vectors is exploited. The least squares cost is introduced, both in its original form, and in an alternative form appropriate to the new least squares recursion. Convergence results are presented.

In the second part of this section, the assumption that \( X_k \) is measured is relaxed and the ad hoc idea of ELS is introduced. The two least squares algorithms are converted, via various assumptions, to a collection of ELS algorithms. Computational and convergence aspects are discussed. The importance of studying these ELS algorithms is both as a motivation for locally convergence RPE algorithms presented in Section 2.4, and as suboptimal computationally efficient approximations to these same RPE algorithms.

2.3.1 Least Squares

In this subsection we consider signal model (2.2.4) given in the previous section, and the idealized estimation task to estimate \( \theta \) given a sequence of observations \( Y_k \), and the state
sequence $X_1, X_2, ..., X_k$. Subsequently, we will consider the case when $X_k$ must be estimated from $Y_k$.

**Off-line**

In a familiar approach, pre-multiplication of (2.2.4) by $X_k$ and some algebraic manipulation leads to the off-line estimate for $\theta$, $\hat{\theta}_k$, based on $k$ data points.

$$\hat{\theta}_k = \left( \sum_{i=1}^{k} X_i X_i' \right)^{-1} \sum_{i=1}^{k} X_i y_i$$ (2.3.1)

where $\hat{\theta}_k$ is an estimate of $\theta$ given $k$ data points. The estimation error is

$$\tilde{\theta}_k = \theta - \hat{\theta}_k = \left( \sum_{i=1}^{k} X_i X_i' \right)^{-1} \sum_{i=1}^{k} X_i w_i.$$ (2.3.2)

By exploiting the idempotent nature of indicator vectors, the above estimates (2.3.1) and (2.3.2) can be written as scalar equations

$$\tilde{\theta}^{(j)}_k = \left( \sum_{i=1}^{k} X_i^{(j)} \right)^{-1} \sum_{i=1}^{k} X_i^{(j)} y_i, \quad \tilde{\theta}^{(j)}_k = \left( \sum_{i=1}^{k} X_i^{(j)} \right)^{-1} \sum_{i=1}^{k} X_i^{(j)} w_i$$ for $j = 1, 2, ..., N.$ (2.3.3)

where $\tilde{\theta}^{(j)}_k$, $\tilde{\theta}^{(j)}_k$ and $X_i^{(j)}$ denote the $j$th element of $\hat{\theta}_k$, $\tilde{\theta}_k$ and $X_i$ respectively.

We are led to the following lemma.

**Lemma 2.3.1** In the above notation, and with $w_k$ a martingale increment with respect to $\mathcal{G}_h$, in that $E[w_k|\mathcal{G}_{h-1}] = 0$, then

$$\lim_{k \to \infty} \hat{\theta}_k \text{ exist a.s.}.$$ (2.3.4)

Moreover, for each $j$ in (2.3.3)

$$\lim_{k \to \infty} \tilde{\theta}^{(j)}_k = \theta^{(j)} \text{ a.s.} \iff \lim_{k \to \infty} \left( \sum_{i=1}^{k} X_i^{(j)} \right)^{-1} = 0.$$ (2.3.5)

**Proof** We proceed first to prove the second result by examining subsequences of the chain on which each state is active. Let $n(k)^j$ denote the number of times state $j$ is active up until time $k$, i.e. $n(k)^j = \sum_{i=1}^{k} X_i^{(j)}$ and let $a(i)^j$ denote the time $k$ at which the state $j$ is active for the $i$th time.
From (2.3.3) we define $W^j_k := \sum_{i=1}^k \left( \sum_{\ell=1}^i X_i^{(j)} \right)^{-1} X_i^{(j)} w_i$ which is a martingale adapted to $\mathcal{G}_k$ since $E[W_{k+1}|\mathcal{G}_k] = W_k$. Note that $W^j_k = \sum_{i=1}^{n(k)^j} \frac{1}{i} w_{a(i)^j}$ by summing only over the subsequence with active $j$, then it follows that $W^j_k$ is bounded in $L_2$ for each $j$, by

$$E \left[ (W^j_k)^2 \right] = E \left[ \sum_{i=1}^{n(k)^j} \frac{1}{i^2} w_{a(i)^j} \right] \leq B \sum_{i=1}^{n(k)^j} \frac{1}{i^2} < \infty$$

where we have used that $E[w_i w_i] = 0$ for all $i \neq \ell$, and that $E[w_i w_i] \leq B$ for all $i$.

For the only if direction of the second lemma result note that under the lemma condition $\lim_{k \to \infty} \left( \sum_{i=1}^k X_i^{(j)} \right)^{-1} = 0$, we have $n(k)^j \to \infty$ as $k \to \infty$. Hence, by the martingale convergence, [64, 59],

$$\lim_{k \to \infty} W^j_k \text{ exist a.s. for each } j.$$ 

Using the Kronecker lemma [53, 64] we have

$$\lim_{k \to \infty} \frac{1}{n(k)^j} \sum_{i=1}^{n(k)^j} w_i = 0 \text{ a.s. for each } j.$$ 

Hence, by rewriting as a summation over $k$, we have

$$\lim_{k \to \infty} \frac{1}{n(k)^j} \sum_{i=1}^k X_i w_i = \lim_{k \to \infty} \tilde{\theta}^{(j)}_k = 0 \text{ a.s. for each } j.$$ 

The result follows. The if direction of the second lemma result follows from noting that $\lim_{k \to \infty} \left( \sum_{i=1}^k X_i^{(j)} \right)^{-1} \neq 0$ implies $n(k)^j$ is finite. Hence,

$$0 < \frac{1}{n(k)^j} \sum_{i=1}^{n(k)^j} w_i < \infty \text{ w.p. } 1.$$ 

The first lemma result now also follows. \hfill \Box

**On-line**

Simple manipulations of (2.3.1) give the recursions

$$\tilde{\theta}_k = \hat{\theta}_{k-1} + P_k x_k [y_k - X_k^{(j)} \hat{\theta}_{k-1}] \text{ and }$$

(2.3.6)
\[ P_{k-1} = P_{k-1}^{-1} + X_k X_k' \]  
\[ P_k = P_{k-1} - P_{k-1} X_k [1 + X_k' P_{k-1} X_k]^{-1} X_k' P_{k-1} \] 

(2.3.7)

(2.3.8)

where \( \hat{\theta}_k \) is an estimate of \( \theta \) after \( k \) data points. Manipulations show that \( \hat{\theta}_k \) minimizes a squares sum index, that is,

\[ \hat{\theta}_k = \arg \min_{\theta} \sum_{i=1}^{k} (y_i - \theta X_i)^2. \]  

(2.3.9)

Now, noting that nonlinear functions on \( X_k \) are linear in \( X_k \) as follows:

\[ f(X_k) = \sum_{i=1}^{N} f(e_i) X_k^{(i)} \]

where \( X_k^{(i)} \) denotes the \( i \)th element of \( X_k \). The above recursion can be rewritten in an alternative form as

\[ \hat{\theta}_k = \hat{\theta}_{k-1} + \sum_{i=1}^{N} P_k e_i [y_k - \hat{\theta}_{k-1} e_i] X_k^{(i)} \]  

(2.3.10)

\[ P_{k-1}^{-1} = P_{k-1}^{-1} + \sum_{i=1}^{N} e_i e_i' X_k^{(i)} \]  

(2.3.11)

\[ P_k = P_{k-1} - \sum_{i=1}^{N} P_{k-1} e_i [1 + e_i' P_{k-1} e_i]^{-1} e_i' P_{k-1} X_k^{(i)} \]  

(2.3.12)

and likewise it can be shown that \( \hat{\theta}_k \) minimizes the linear index

\[ \hat{\theta}_k = \arg \min_{\theta} \sum_{i=1}^{k} \sum_{j=1}^{N} (y_i - \theta e_j)^2 X_i^{(j)} . \]  

(2.3.13)

**Remarks**

1. The condition that \( \left( \sum_{i=1}^{k} X_i^{(i)} \right)^{-1} \rightarrow 0 \) as \( k \rightarrow \infty \) in Lemma 2.3.1 is an excitation condition. Also, it is possible to show, but is not done here that the rate of convergence to zero of \( \hat{\theta}_k \hat{\theta}_k' \), is as \( 1/k^+ \). See also Stenby [75].

2. To reduce the number of calculations, equations (2.3.7) and (2.3.11) can be replaced by a stochastic approximation given by

\[ P_{k}^{-1} = k E[X_k X_k'] = k \text{diag}(E[X_k]) \]

where \( E[X_k] \) is a vector of the a prior probabilities of being in each state. \( E[X_k] \) is given by the normalized eigenvector of \( A \) corresponding to the eigenvalue of value \( 1 \).

We now proceed to consider the case when the state sequence is unknown.
2.3.2 Extended Least Squares

This subsection considers the estimation task when the state sequence is not known, and is presented in a manner paralleling the previous section. To produce estimates when the state sequence is unknown, the ad hoc idea of extended least squares is to use estimates of states in lieu of actual states $X_k$, in a least squares implementation, see [72]. We will show in the following discussion that substitution of state estimates, into the two alternative forms of the LS algorithm shown in the previous section, results in two different ELS algorithms.

In linear estimation it is usual to use one-step-ahead predictions of the states so that the observation noise remains white, at least when the predictions are optimal in a least squares sense. However in our case, the one-step-ahead predictions of Markov chains can be far from optimal, particularly when the active state changes. This property of HMM highlights that there are differences between standard linear estimation theory and HMM parameter estimation problem. Initially in this section we proceed by using one-step-ahead predictions to ensure that the observation noise remains white, but this requirement is relaxed towards the end of this section.

Let $\hat{X}_{k|k-1,\theta_{k-1}}$ denote the conditional filter estimate based on observations and model estimates, i.e.,

$$\hat{X}_{k|k-1,\theta_{k-1}} := E[X_k|Y_{k-1}, \hat{\Theta}_{k-1}]$$

where $\hat{\Theta}_{k-1} := \{\hat{\theta}_1, \ldots, \hat{\theta}_{k-1}\}$ and the below recursion is used to generate the one-step-ahead predictions

$$\hat{X}_{k+1|k,\theta_k^*} = N_k AB(y_k, \hat{\theta}_k^*) \hat{X}_{k|k-1,\theta_{k-1}}^*.$$  \hspace{1cm} (2.3.14)

Off-line

Using the ad hoc idea of replacing states by one-step-ahead predictions, the extended least squares version of the off-line least squares algorithms (2.3.3) is

$$\hat{\theta}_k^j = \left[ \sum_{i=1}^{k} \hat{X}_{i|j-1,\theta_{i-1}}^{(j)} \right]^{-1} \left[ \sum_{i=1}^{k} \hat{X}_{i|j-1,\theta_{i-1}}^{(j)} y_i \right], \text{ for } j = 1, 2, \ldots, N \hspace{1cm} (2.3.15)$$

where $\hat{\theta}_k$ is the estimate of $\theta$ on $k$ points of data and $\hat{X}_{i|j-1,\theta_{i-1}} = E[X_i|Y_j, \hat{\Theta}_{i-1}]$. No convergence analysis is attempted here for the off-line ELS algorithm.
§2.3 Least Squares and Extended Least Squares

On-line

From the least squares recursion (2.3.6)-(2.3.8), substituting one-step-ahead predictions gives the recursions:

\[
\hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \hat{X}_{k|k-1, \hat{\theta}_{k-1}} [y_k - \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \hat{\theta}_{k-1}]
\]  
(2.3.16)

\[
\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \text{ or}
\]

\[
\hat{P}_k = \hat{P}_{k-1} - \sum_{i=1}^{N} \hat{P}_{k-1} \hat{X}_{k|k-1, \hat{\theta}_{k-1}} [y_k - \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \hat{P}_{k-1}]^{-1} \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \hat{P}_{k-1}.
\]  
(2.3.18)

Likewise, from (2.3.10) we construct the ELS recursion:

\[
\hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \sum_{i=1}^{N} e_i [y_k - \hat{\theta}_{k-1} e_i]
\]  
(2.3.19)

\[
\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \text{ or}
\]

\[
\hat{P}_k = \hat{P}_{k-1} - \sum_{i=1}^{N} \hat{P}_{k-1} e_i e_i \hat{P}_{k-1} [y_k - \hat{\theta}_{k-1} e_i]^{-1} \hat{X}_{k|k-1, \hat{\theta}_{k-1}}
\]  
(2.3.21)

where \(\hat{X}_{k|k-1, \hat{\theta}_{k-1}}\) is the diagonal matrix with \(\hat{X}_{k|k-1, \hat{\theta}_{k-1}}\) on its diagonal. Note that for (2.3.20) if \(\hat{P}_0\) is diagonal then \(\hat{P}_k\) will be diagonal for all \(k\). Hence equations (2.3.20) and (2.3.21) can be explicitly written as scalar equations

\[
\hat{p}_{i,k}^{-1} = \hat{p}_{i,k-1}^{-1} + \hat{X}_{k|k-1, \hat{\theta}_{k-1}} \text{ or} \hat{p}_{i,k} = \hat{p}_{i,k-1} - \frac{\hat{p}_{i,k-1} \hat{X}_{k|k-1, \hat{\theta}_{k-1}}}{1 + \hat{p}_{i,k-1} \hat{X}_{k|k-1, \hat{\theta}_{k-1}}}
\]  
(2.3.22)

where \(\hat{P}_k = \text{diag}([\hat{p}_{i,k}, ..., \hat{p}_{i,k}, ..., \hat{p}_{N,k}])\).

From a computational point of view, the diagonal nature of \(\hat{P}_k\) in (2.3.19) means that the (2.3.19)-(2.3.21) is computationally more attractive for obtaining the \(\theta\) estimates then (2.3.16)-(2.3.18), being of order \(N\) rather than \(N^2\) in complexity.

However, (2.3.16)-(2.3.18) are not completely satisfactory because the estimates produced are biased. To see this we now examine the convergence properties of (2.3.16)-(2.3.18) for the idealized case when

\[
\hat{X}_{k|k-1, \hat{\theta}_{k-1}} = \hat{X}_{k|k-1, \hat{\theta}} = E[\hat{X}_k | Y_{k-1}, \hat{\theta}],
\]  
(2.23.23)

**Lemma 2.3.2** Consider the ELS scheme (2.3.19)-(2.3.21) in the idealized case, when (2.23.23) holds. Then as \(k \to \infty\)

\[
\hat{\theta}_k \to \hat{P}_k \left[ \sum_{j=1}^{k} \hat{X}_{j|j-1, \theta} X_{j} \right] \theta \ a.s.
\]
\[ \hat{\theta}_k = \hat{P}_k \left[ \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} X_j \right] \theta + \hat{P}_k \left[ \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} w_j \right] \]  

(2.3.26)

Now

\[ \left[ \frac{1}{k} \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} \right]^{-1} \left[ \frac{1}{k} \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} w_j \right] \to 0 \text{ a.s.} \]  

(2.3.27)

since, each element of the second term can be shown to go to zero using martingale convergence results and the Kronecker Lemma, as in the proof of Lemma 2.3.1. Also observe that

\[ \hat{P}_k \left[ \sum_{j=1}^{k} \hat{X}_{jj-1,\theta}^2 \right] \leq I \]  

(2.3.28)

since \(0 \leq \hat{X}^2 \leq \hat{X} \leq 1\). Clearly (2.3.25)-(2.3.27) yield the first half of (2.3.24) as claimed.

Now observe that since \(E[X_j - \hat{X}_{jj-1,\theta} Y_{j-1}, \theta] = 0\), then likewise

\[ \hat{P}_k \left[ \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} (X_j - \hat{X}_{jj-1,\theta} Y_j) \right] \to 0 \text{ a.s.} \]  

(2.3.29)

and yield the second half of (2.3.24).

Now from (2.3.16) and (2.3.19), simple manipulations lead to

\[ \hat{\theta}^* = \hat{P}_k^* \left[ \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} y_j \right] \]  

\[ = \hat{P}_k^* \hat{P}_k^{-1} \theta_k \]  

(2.3.30)

and the result (2.3.25) follows from (2.3.26), under the excitation assumption that \(\hat{P}_k^* \to 0\) so that \(||\hat{P}_k^* \hat{P}_k^{-1}||\) is bounded above and \(\hat{P}_k^* \sum_{j=1}^{k} \hat{X}_{jj-1,\theta} w_j \to 0 \text{ a.s. as } k \to \infty\).  

\[ \Box \]

Remarks
1. The lemma result (2.3.25) holds without the excitation condition on \( \hat{P}_k \), which incidentally assures a convergence rate of \( 1/k^{1/2} \) but more advanced theory such as in [75] is required.

2. Lemma 2.3.2 demonstrates that the scheme (2.3.19)-(2.3.21) leads to biased estimates whereas the scheme (2.3.16)-(2.3.18) does not. Hence, because both schemes are of similar complexity, it seems that the scheme (2.3.16)-(2.3.18) should be used in preference. See the simulations section for a demonstration of the bias.

3. The complete ELS convergence analysis of (2.3.16)-(2.3.18) when (2.3.23) does not hold is virtually identical to that given [40], and is not repeated here. Suffice it to say, a key sufficient condition for (2.3.23)-(2.3.25) to hold asymptotically is that a certain passivity condition holds. That is, the system driven by \( \tilde{\theta}_k \hat{X}_k \) and with output \( \frac{1}{2} \tilde{\theta}_k \hat{X}_k + \theta \hat{X}_k \) be strictly passive (here \( \tilde{\theta} = \theta - \hat{\theta} \) and \( \hat{X} = X - \hat{X} \)). This system in the HMM case is nonlinear and is sample path dependent, so further explorations along this line seems pointless.

4. Consider also a hybrid version of (2.3.16)-(2.3.18)

\[
\tilde{\theta}_k = \tilde{\theta}_{k-1} + \hat{P}_k \tilde{X}_{k|k-1}, \tilde{\theta}_{k-1} \left[ y_k - \hat{X}_{k|k-1}, \tilde{\theta}_{k-1} \right] \tag{2.3.31}
\]

with (2.3.20) and (2.3.21) holding. We do not study (2.3.31) further here.

5. To further reduce the number of calculations required to estimate \( \theta \), equation (2.3.20) can be replaced by a stochastic approximation given by

\[
P_k^{-1} = kE[\tilde{X}_{k|k-1}, \tilde{\theta}_{k-1}] = kdiag(E[X_k]).
\]

A Posteriori Extended Least Squares

It is the nature of HMMs that the one-step-ahead predictions of the state can be far from optimal, particularly when the active state changes.

Hence, here we consider a ELS algorithm based on filtered estimates rather than one-step-ahead predictions. Consider a modified version of the (2.3.19)-(2.3.21) scheme.
\[ \hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \hat{X}_{k|k;\hat{\theta}_{k-1}} \sum_{i=1}^{N} e_i[y_k - \hat{\theta}_{k-1} e_i] \]  
(2.3.32)

\[ \hat{P}_{k-1} = \hat{P}_{k-1}^i + \hat{X}_{k|k;\hat{\theta}_{k-1}}, \quad \text{or} \]  
(2.3.33)

\[ \hat{P}_k = \hat{P}_{k-1} - \sum_{i=1}^{N} \hat{P}_{k-1} e_i e_i^T \hat{P}_{k-1} [1 + e_i \hat{P}_{k-1} e_i]^{-1} \hat{X}_{k|k;\hat{\theta}_{k-1}}. \]  
(2.3.34)

**Remarks**

1. The recursion (2.3.32)-(2.3.34) is computationally efficient because \( \hat{P}_{k-1} \) can be forced to be diagonal.

2. Unlike the recursions (2.3.19)-(2.3.21), the recursion (2.3.32)-(2.3.34) produces consistent results in simulations, see Section 2.5.

This ELS recursion is the most attractive of the algorithms presented in this section, however, no martingale convergence analysis is available for the (2.3.32)-(2.3.34) scheme at present since the error term \( X_k - \hat{X}_{k|k} \) is not a martingale increment. Rather than proceed with a further analysis of this a posteriori ELS scheme, we proceed to look at RPE algorithms which are mildly more complicated but have a more complete theory.

**2.4 Recursive Prediction Algorithms**

There exists mature theory for recursive estimation or identification of continuous discrete-time models based on the minimization of the prediction error costs, see [51]. This theory provides asymptotic quadratic convergent algorithms (admittedly local) for linear and non-linear models.

In this section we proceed by applying this mature theory, in order to obtain asymptotic convergence algorithms which generalize the ELS schemes of the last section. First, we re-introduce the two cost functions from the least squares discussion to replace the usual prediction error cost. These cost functions are our criteria for estimation of \( \theta \). We present the RPE algorithm which minimizes each of these cost functions.

Following on from this we present an RPE algorithm corresponding to the attractive a posteriori extended least squares algorithm (2.3.32)-(2.3.34).
Prediction Error Cost Functions

First, we present the RPE schemes corresponding to the computational efficient one-step-ahead prediction based ELS schemes, (2.3.31) and (2.3.19)-(2.3.21).

Consider the error cost functions

\[
V_k(\theta) := \frac{1}{2} \sum_{i=2}^{k} \left(y_i - \theta'\bar{X}_{i|k-1,\tilde{\theta}_{k-1}} \right)^2 \quad \text{and} \quad (2.4.1)
\]

\[
\tilde{V}_k(\theta) := \frac{1}{2} \sum_{i=2}^{k} \sum_{j=1}^{N} \left(y_i - \theta' e_j \right)^2 \tilde{X}_{i|k-1,\tilde{\theta}_{k-1}}, \quad (2.4.2)
\]

The following RPE recursion minimizes locally the index in (2.4.1), see Lemma 2.4.1, and generalizes the ELS recursion (2.3.31):

\[
\begin{align*}
\tilde{\theta}_k^i &= \tilde{\theta}_{k-1}^i + \tilde{P}_k^i \psi_k \bar{\theta}_{k-1}^i \\
\tilde{P}_k^i &= \tilde{P}_{k-1}^i + \tilde{X}_{k|k-1,\tilde{\theta}_{k-1}} \quad (2.4.3)
\end{align*}
\]

where

\[
\psi_k|k-1,\tilde{\theta}_{k-1}^i := \left. \frac{d}{d\theta} (y_k - \bar{X}_{k|k-1,\tilde{\theta}_{k-1}})^2 \right|_{\theta = \tilde{\theta}_{k-1}^i},
\]

\(\tilde{P}_{k}^i\) is an approximation for the second derivative of \(V_k(\theta)\), and from (2.3.14)

\[
\tilde{X}_{k+1|k,\tilde{\theta}_{k}^i} = N_{k} A B (y_{k}, \tilde{\theta}_{k}^i) \tilde{X}_{k|k-1,\tilde{\theta}_{k-1}}.
\]

The RPE recursion which minimizes locally the index (2.4.2), see Lemma 2.4.1, and generalizes the ELS recursion (2.3.19)-(2.3.21) is

\[
\begin{align*}
\tilde{\theta}_k &= \tilde{\theta}_{k-1} + \tilde{P}_k \kappa_{k|k-1,\tilde{\theta}_{k-1}} \\
\tilde{P}_{k} &= \tilde{P}_{k-1} + \tilde{X}_{k|k-1,\tilde{\theta}_{k-1}} \quad (2.4.4)
\end{align*}
\]

where, with \(\kappa^{(i)}\) denoting the \(i\)th element of \(\kappa\),

\[
\kappa_{k|k-1,\tilde{\theta}_{k-1}}^{(i)} = \tilde{X}_{k|k-1,\tilde{\theta}_{k-1}}^{(i)} [y_k - \tilde{\theta}_{k-1}^{(i)}] - \sum_{j=1}^{N} \left. \frac{d\tilde{X}_{k|k-1,\tilde{\theta}_{k-1},\theta}^{(j)}}{d\theta^{(i)}} \right|_{\theta = \tilde{\theta}_{k-1}} [y_k - \tilde{\theta}_{k-1}^{(j)}]^2.
\]

Convergence of both these RPE algorithms can be established by a conventional ODE analysis [51]. Since the state estimates \(\tilde{X}_{i|k-1,\tilde{\theta}_{k-1}}\) and \(\tilde{X}_{i|k-1,\tilde{\theta}_{k-1}}\) are of necessity bounded, a projection into a stability domain as required in [13] is implicit here.
Actually, the ODE analysis requires that the filter generating $\hat{X}_{k|k-1,\hat{h}_{k-1}}$ be exponentially stable. This exponential stability in the sense that initial conditions are forgotten exponentially, is established in [8] for the $N = 2$ case, and is known to hold more generally under reasonable conditions not spelt out here.

To demonstrate convergence of (2.4.3) and (2.4.4), first, let us define for (2.4.3) and (2.4.4), respectively, and arbitrary $\theta$,

$$ f(\theta, k) = E[\psi_{k|\theta \epsilon_{k|k-1,\theta}}], \text{ or } E[\kappa_{k|k-1,\theta}] \quad \text{and} \quad (2.4.5) $$

$$ G(\theta, k) = E[\hat{X}_{k|k-1,\theta}]. \quad (2.4.6) $$

The following lemma now holds

**Lemma 2.4.1** The recursions (2.4.3) and (2.4.4) will converge a.s. to the set

$\tilde{\mathcal{D}}_0 = \{\theta|\lim_{k \to \infty} E[f(\theta, k)] = 0\}$, moreover, under the excitation condition $\hat{F}_k (\text{ or } \hat{F}_k^\dagger) \to 0$ as $1/k$ then convergence of $\tilde{\theta}_k (\text{ or } \tilde{\theta}_k^\dagger)$ is at the rate $1/k^+. $

**Proof** The ODEs associated with (2.4.3) and (2.4.4) for fixed $k$, under (2.4.5) and (2.4.6), are

$$ \frac{d}{d\tau} \theta(\tau, k) = R^{-1}(\tau, k) f(\theta(\tau, k), k) $$

$$ \frac{d}{d\tau} \tau(\tau, k) = G(\theta(\tau, k), k), \quad R_0(k) \geq \delta I. \quad (2.4.7) $$

Now Lyapunov functions for (2.4.7) under (2.4.5) and (2.4.6), are

$$ \tilde{W}(\tau, k) = E[(y_k - \theta \hat{X}_{k|k-1,\theta})^2], \quad \text{or} \quad \tilde{W}(\tau, k) = E[\sum_{j=1}^{N} (y_k - \theta \epsilon_{j})^2 \hat{X}_{k|k-1,\theta}] \quad (2.4.8) $$

so that

$$ \frac{d}{d\tau} \tilde{W}(\tau, k) = \frac{d\tilde{W}(\tau, k)}{d\theta(\tau, k)} \frac{d\theta(\tau, k)}{d\tau} $$

$$ = -f'(\theta(\tau, k), k) R_0^{-1}(\tau) f(\theta(\tau, k), k). \quad (2.4.9) $$

Thus $\tilde{W}(\tau, k)$ converges for all $k$ and $\tau \to \infty$, and $\theta(\tau, k)$ converges to the set $\{\theta|E[f(\theta, k)] = 0\}$

Applying the ODE theory of Ljung [51], the various regularity conditions are satisfied here and the first result claimed follows.
Observe from (2.4.9) that if \( R_r(k) \) is of the order \( 1/k \), as under suitable excitation, then \( f(\theta(k), k) \) converges to zero as \( 1/k^4 \). Since asymptotically the stochastic difference equation behaves as the ODE, then rates of convergence translate across under the scaling of the theory.

This leads to the convergence rate result of the lemma.

\[ \square \]

**Remarks**

1. The RPE schemes are mildly more sophisticated than the ad hoc one-step-ahead prediction based ELS schemes of the previous section. For this reason we have kept the same ELS notation to assist in seeing the similarities and differences.

2. For a RPE version of (2.3.16) see [13].

3. The equations (2.4.3) and (2.4.4) do not result from standard RPE theory. The search direction has been changed, so that \( P_k^{-1} \) is diagonal, but the scheme still provides quadratic convergence.

4. A complete and precise theory on convergence rates is not given in the above results, this being beyond the scope of this chapter.

5. To reduce the number of calculations, the second half of equations (2.4.3) and (2.4.4) can be replaced by a stochastic approximation given by

\[ P_k^{-1} = kE[X_k X_k^\prime] = k\text{diag}(E[X_k]). \]

Convergence can still be proven with a slight modification of lemma 2.4.1.

**A Posterior Weighted RPE Scheme**

To generalize the ELS algorithm based on filtered estimates (2.3.32)-(2.3.34) rather than one-step-ahead predictions, we consider RPE schemes based on filtered estimates. To do so consider the following cost function.

\[ \hat{V}_k(\theta) := \frac{1}{2} \sum_{i=2}^{k} \sum_{j=1}^{N} (y_i - \theta e_j)^2 \hat{X}_{i,j} \hat{X}_{i,j}^\prime, i, \theta \]

which gives the updates equations given by
\[
\begin{align*}
\hat{\theta}_k &= \hat{\theta}_{k-1} + \hat{P}_k \kappa_{k|k, \hat{\theta}_{k-1}} \\
\hat{P}_k^{-1} &= \hat{P}_{k-1}^{-1} + \hat{X}_k|k, \hat{\theta}_{k-1}.
\end{align*}
\tag{2.4.11}
\]

with \(\kappa^{(i)}\) the \(i\)th element of \(\kappa\) defined from

\[
\kappa^{(i)}_{k|k, \hat{\theta}_{k-1}} = -\hat{X}^{(i)}_{k|k, \hat{\theta}_{k-1}} [y_k - \hat{\theta}^{(i)}_{k-1}] + \sum_{j=1}^{N} \left. \frac{d \hat{X}^{(j)}_{k|k, \hat{\theta}_{k-1}, \theta}}{d \theta^{(i)}} \right|_{\theta = \hat{\theta}_{k-1}} [y_k - \hat{\theta}^{(j)}_{k-1}]^2 \tag{2.4.12}
\]

and

\[
\hat{X}_k|k, \hat{\theta}_k = N_k B(y_k, \hat{\theta}_k) A \hat{X}_{k-1|k-1, \hat{\theta}_{k-1}}. \tag{2.4.13}
\]

**Lemma 2.4.2** The recursion (2.4.11) will converge a.s. to the set

\(\bar{D}_0 = \{\theta | \lim_{k \to \infty} E[f(\theta, k)] = 0\}\), moreover, under the excitation condition \(\hat{P}_k (or \hat{P}^+_k) \to 0\) as \(1/k\) then convergence of \(\hat{\theta}_k (or \hat{\theta}^+_k)\) is at the rate \(1/k^\frac{1}{2}\).

**Proof** Same as for Lemma 2.4.1 with \(f(\theta, k) = E[\kappa_{k|k, \theta}]\) and \(G(\theta, k) = E[\hat{X}_{k|k, \theta}]\) and using the Lyapunov function

\[
\tilde{W}(\tau, k) = E[\sum_{j=1}^{N} (y_k - \theta^* e_j)^2 \hat{X}^{(j)}_{k|k, \theta}].
\]

\(\square\)

**Remarks**

1. The cost function (2.4.10) is the sum of the predicted error of being in each state, weighted by the estimated probability of being in each state, which from (2.2.6) is \(\hat{X}_{k|k, \theta}\)

2. The similarity of form between (2.4.11) and (2.3.32) suggest that the recursions (2.3.32) are valid at least as approximations for (2.4.11), for which convergence has been shown.

3. Again to reduce the number of calculations, the second half of equation (2.4.11) can be
replaced by an stochastic approximation given by

\[ P_k^{-1} = kE[X_kX'_k] = k\text{diag}(E[X_k]) \]

and the convergence proof holds.

\section*{2.5 Implementation Considerations and Simulations Results}

This section has two parts. In the first part, issues concerning implementation of algorithms for estimating \( \theta \) are presented. The discussion is general in nature and in fact applies to any of the algorithms presented in this chapter, and others in the literature of this field.

In the second part, simulation studies of the various algorithms present in this chapter are presented. The simulation studies are not presented to verify the algorithms but only to highlight the various properties of these algorithms. The highlight properties include: convergence, convergence rates, bias, and the importance of the issues introduced in the first part of this section.

\subsection*{2.5.1 Implementation Considerations}

The following were considered when implementing the schemes presented in the preceding chapters.

\textbf{Transients}

One reason for studying both ELS and RPE schemes in the same chapter is that it appears a good approach to use them in combination in an actual implementation. The extra gradient terms used in the RPE schemes do not assist during the transient period where the dominant error is due to initialization rather than the noise, however, these terms do aid convergence subsequently. Thus it is a reasonable practise to use an ELS scheme initially and change to an RPE scheme once the transient has decayed significantly.

\textbf{Step Sequence}

Step size adjustments can be made for improved transient performance for iterative schemes, and indeed \( \hat{P}_k = \frac{1}{\gamma_k} \hat{R}_k^{-1} \) can be replaced by \( \gamma_k \hat{R}_k^{-1} \) for arbitrary \( \gamma_k \) satisfying \( \sum_{k=1}^{\infty} \gamma_k = \)
\[ \sum_{k=1}^{\infty} \gamma_k^2 < \infty \] and the ODE analysis still applies. Further details are omitted here, see Ljung [51].

**Markov State Errors**

The time-varying variance of the state estimate, given by

\[
\Sigma_k = E[(X_k - \bar{X}_k | k-1, \Theta_{k-1}')(X_k - \bar{X}_k | k-1, \Theta_{k-1}')']
\]

\[
= \bar{X}_k | k-1, \Theta_{k-1} - \bar{X}_k | k-1, \Theta_{k-1}, \bar{X}_k' | k-1, \Theta_{k-1} \geq 0 \quad (2.5.1)
\]

can be used in the recursive equations to “discount” time instants for which the Markov state is known with less certainty. If the variance of the state estimate is denoted \( \sigma_k^2 = \Theta_k \Sigma_k \Theta_k' \), then the modified update equation, according to standard Kalman filter theory, becomes

\[
\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \frac{1}{\sigma_w^2 + \sigma_k^2} \bar{X}_k | k-1, \Theta_{k-1} \quad (2.5.2)
\]

Also, in (2.3.19), (2.4.3) and (2.4.4) \( \hat{P}_k \) is replaced by \( \frac{1}{\sigma_w^2 + \sigma_k^2} \hat{P}_k \). Corresponding modifications apply to (2.3.16), (2.3.17) and (2.3.31).

**Parameter Estimation Errors**

Similarly, the variance of the parameter estimates, approximated by \( \hat{P}_k \), can be used to modify the variance used in (2.2.5) to estimate the Markov states.

\[
\sigma_w^{(m)} = \sigma_w^2 + \bar{X}_k' \hat{P}_k \bar{X}_k. \quad (2.5.3)
\]

Actually, in practice it makes sense to limit the magnitude of additive term to \( \sigma_w^2 \), to say \( \sigma_w^{(m)} \), because of the approximations involved. That is,

\[
\sigma_w^{(m)} = \sigma_w^2 + \min \{ \sigma_w^2, \bar{X}_k' \hat{P}_k \bar{X}_k \}. \quad (2.5.4)
\]

**Polyak Acceleration**

The increased step size and averaging used by Collings [13] is suggested by Polyak [65] as a technique to speed convergence. The Polyak increased step size has been found, in some cases, to aid convergence from poor initial estimates and in high noise.
Time-varying Tracking

It is possible to modify the estimation schemes presented in this chapter to allowing tracking of time varying parameters by introducing a forgetting factor, see Ljung [51]. A forgetting factor, \( \lambda \), is introduced by modifying the second equations of (2.4.3) or (2.4.11), or the corresponding ELS schemes, to give

\[
\hat{P}_k^{-1} = \lambda \hat{P}_{k-1}^{-1} + \hat{X}_{k|k-1, \theta_{k-1}}
\]

(2.5.5)

where typically \( \lambda \leq 1 \).

This modification was also found to improve convergence in very high noise simulations.

2.5.2 Simulations

We present results of simulation examples using computer generated finite discrete state Markov chains. The results presented in the following simulation plots were found to be representative examples of hundreds of simulation runs. We concentrated our efforts on the new algorithms that appear interesting. For example, the most tested algorithms are (2.3.16), (2.3.32), (2.3.31), and (2.4.11). The least tested algorithms are (2.3.6) and (2.3.19).

Convergence using \( \hat{X}_{k|k-1, \theta} \)

A two-state Markov chain embedded in white Gaussian noise (WGN) is generated with parameter values \( a_{ii} = 0.85, a_{ij} = (1 - a_{ii}) \) for \( i \neq j \), \( \theta = [3, 6]' \), \( \sigma_w = 0.5 \). The state sequence, \( \hat{X}_{k|k-1, \theta} \), is estimated (ie, assuming knowledge of \( \theta \), so that (2.3.23) holds) and the state values \( \theta \) are estimated from \( \hat{X}_{k|k-1, \theta} \). Each of the following schemes is used to estimate the parameters \( \theta \):

- A least squares algorithm (2.3.6) with \( X_k \) known;
- Original ELS algorithm (2.3.16), using \( \hat{X}_{k|k-1, \theta} \);
- RPE from [6], using \( \hat{X}_{k|k-1, \theta} \);
- A posterior ELS algorithm (2.3.32), using \( \hat{X}_{k|k-1, \theta} \);
- Hybrid ELS algorithm (2.3.31), using \( \hat{X}_{k|k-1, \theta} \).

Figure 2.1 shows an empirical comparison of the rates of convergence. This figure shows the convergence of various schemes to one of the parameters.
Comparison of $\hat{X}_{k|k-1,\delta_{k-1}}$ with $\hat{X}_{k|k-1,\delta}$

A two-state Markov chain embedded in WGN is generate from which $\hat{X}_{k|k-1,\delta_{k-1}}$ and $\hat{X}_{k|k-1,\delta}$ are estimated using (2.3.32). Figure 2.2 shows the difference between the estimates over time. It appears in this simulation that asymptotically condition (2.3.23) holds. Note that the average over 100 points is used in this figure to reduce the amount of information presented.

Bias of estimation

To verify remark 2 made in subsection 2.3.2, that a bias is indeed introduced by (2.3.19) a two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.75$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [3, 6]'$, $\sigma_w = 1$. The state values are estimated using (2.3.16) and (2.3.19) and noting that the state estimates $\hat{X}_{k|k-1,\delta}$ are used. The estimated parameter from the schemes are: $\hat{\theta}_{2.3.16} \approx [3.01, 5.86]'$ and $\hat{\theta}_{2.3.19} \approx [4.23, 4.76]'$. The estimates obtained by (2.3.19) are indeed biased away from the true parameter values.

Convergence rate comparison

A two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.70$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, $\sigma_w = 1$. The state values are estimated using the following schemes:

- A least squares algorithm (2.3.6) with $X_k$ known;
- Original ELS algorithm (2.3.16), using $\hat{X}_{k|k-1,\delta_{k-1}}$;
- RPE of [13], using $\hat{X}_{k|k-1,\delta_{k-1}}$;
- A posterior ELS algorithm (2.3.32), using $\hat{X}_{k|k-1,\delta_{k-1}}$;
- RPE algorithm (2.4.11), using $\hat{X}_{k|k-1,\delta_{k-1}}$.

Figure 2.3 shows an empirical comparison of the rates of convergence. This figure shows the convergence of these schemes to one of the parameters. This suggests that our $O(N)$ schemes converge at approximately the same rate.

Comparison with existing the RPE scheme

A two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.75$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, $\sigma_w = 1$. The state values are estimated from
the chain using (2.3.32) and the $O(N^2)$ RPE scheme presented in [13]. Figure 2.4 shows an empirical comparison of the rates of convergence to one of the state values. This figure shows that similar rates of convergence are achieved by our new schemes with computational requirements of $O(N)$ and the existing RPE scheme [13] with requirements $O(N^2)$.

**Stochastic approximation**

A two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.80$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, $\sigma_w = 1$. The state values are estimated from the chain using (2.3.32) with and without the approximation described in remark 5. Figure 2.5 shows that convergence of the scheme was not adversely affected by the approximation.

**Fast Markov chains**

A three-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.65$, $a_{ij} = (1 - a_{ii})/2$ for $i \neq j$, $\theta = [0, 1, 2, 3]'$, $\sigma_w = 1$. The state values are estimated using the recursive schemes (ELSM := (2.3.32) recursion) and (RPE1 := (2.4.3) recursion). The ELSM converges to the correct values while the RPE1 scheme does not. This simulation demonstrates that the RPE1 recursions do not estimate low inertia Markov chains well. In this simulation and others involving fast Markov chains the ELSM recursions were found to perform better than the RPE1 recursions. Here, $a_{ii} = 0.65$ implies short times in each state.

This is a significant result. The (2.3.32) recursion is the only recursion we have studied which can handle fast chains effectively. These fast chains are known to appear often in actually applications such as the demodulation of coded QAM signals which is under study in a later chapter.

**A six-state example**

A six-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.95$, $a_{ij} = (1 - a_{ii})/2$ for $i \neq j$, $\theta = [1, 2, 3, 4, 5, 6]'$, $\sigma_w = 1$. The state values are estimated using the RPE1 recursions. Figure 2.6 show the parameter convergence of the RPE1 recursions. Note, in these simulations the Polyak increased step size is used to allow convergence from poor initial estimates.
High noise

A two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.9$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [-1, 1]$, $\sigma_w = 40$. The states are estimated using two methods. Firstly, Figure 2.7 shows the convergence of parameters using the increased step sequence $1/\sqrt{n}$ and averaging over 1000 points. Secondly, Figure 2.8 shows the convergence of parameters using a scheme modified to track time-varying parameters with a forgetting factor, $\lambda = 0.995$. Both figures shows that it is possible to estimates state values in very high noise environments.

Variance corrections

From (2.5.2) and (2.5.3) we can see that the correction factors are only going to have an effect when not dominated by $\sigma^2_w$. These variance correction factors only improve the performance of these schemes during initial transients or in low noise situations.

For example, a two-state Markov chain embedded in WGN is generated with parameter values $a_{ii} = 0.85$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]$, $\sigma_w = 0.00001$. The states are estimated using (2.3.32) scheme both with and without the variance correction factors given by (2.5.2) and (2.5.3). Figure 2.9 shows that the scheme without correction factors did not converge while the scheme using the correction factors does. In low measurement noise situations the estimation noise dominates the measurements noise and it must be included to achieve reasonable results. In high measurement noise situations the relative contribution of the estimation noise is negligible, and need not be considered.

Summary

A variety of schemes presented in this chapter were shown, at least in the simulation examples presented, to provide competitive convergence performance in comparison with previous work present in [13]. In addition to this, several of the implementation issues raised in the previous subsection were examined.

It is noted that in the examples shown, the scheme (2.3.32)-(2.3.34) provides good convergence performance in situations involving high noise and/or fast Markov chains which exposed the limitations of the previous RPE algorithm [13]. More extensive simulation studies would be need to establish the reliability of this scheme.
2.6 Conclusions

In this chapter we have proposed new on-line parameter estimation schemes for HMMs based on extended least squares and recursive prediction error methods. The transition probabilities between states are assumed known, but the state values between which the noise-free measurements switch are learnt in time. These new schemes exploit the idempotent property of the signal model states, noting that care must be taken for the ELS schemes to avoid bias. We present illustrative simulation examples of the schemes in a variety of conditions and highlight the similarity and the difference between the performance of these schemes and compare them with the existing RPE scheme for parameter estimation. The algorithms presented have computational complexity $O(N)$, yet perform as well asymptotically as earlier schemes proposed of $O(N^2)$.

The a posteriori ELS and a posteriori weighted RPE scheme, which exploit filtered state estimates rather than prediction estimates, appear to be the most attractive for application purposes. These a posteriori schemes have been also found to be consistent and thus attractive in signalling environments that include low inertia HMMs which could not be handled well by earlier algorithms [13].

![Graph](chart.png)

*Figure 2.1: Empirical comparison of proposed schemes under assumption (3.23).*
Figure 2.2: Convergence to idealized state estimates.

Figure 2.3: Empirical comparison of proposed schemes.
Figure 2.4: Empirical comparison of the ELS scheme (3.32) and the RPE scheme of [6].

Figure 2.5: Effect of stochastic approximation on convergence.
Figure 2.6: Parameter convergence of a six-state chain example.

Figure 2.7: Parameter convergence in very high noise using Polyak acceleration. Averaging is performed over 1000 points.
Figure 2.8: Parameter convergence in very high noise using forgetting factors.

Figure 2.9: The effect of variance correction factors.
Chapter 3

Estimation of the Transition Probabilities

3.1 Introduction

In the previous chapter new algorithms were proposed for estimating the state output mapping $C$ via extended least squares (ELS) and RPE techniques. These algorithms exploit the discrete state structure of HMMs in ways for which there is no parallel in standard state space model identification. The computational effort of the algorithms presented in Chapter 2 is also less than that for the algorithm presented in [13]. In this chapter, we exploit and build on the ideas of Chapter 2 to produce algorithms for estimating the stochastic matrix $A$ with similar improvements in computational requirements and without computational difficulties as the noise level decreases.

The key contribution of this chapter is a new recursive algorithm based on a state prediction error cost function, rather than based on the output prediction error cost function used in [13]. The recursive state prediction error (RSPE) algorithm proposed here is shown to minimize the state prediction error cost and has less computational requirements than the scheme presented in [13]. An ELS algorithm is also proposed which requires computational effort of only $O(N^2)$ each time instant, compared with the $O(N^4)$ required for the RSPE and RPE schemes. Complete ordinary differential equation (ODE) convergence analysis is presented for the RSPE algorithm but convergence analysis for the proposed ELS algorithm has not been completed. We also show that the proposed RSPE algorithm evanesces to the ELS algorithm and indeed to the least squares (LS) algorithm as the signal to noise ratio increases.

49
A second contribution of this chapter is a scheme which allows simultaneous estimation of the state output mappings $C$ and the state transition probability matrix $A$. The proposed scheme requires less computational effort than the simultaneous estimation scheme presented in [13], but still requires $O(N^4)$ calculations per time instant.

This chapter is organized as follows: In Section 3.2 the signal model, conditional state estimates and a parametrized information state model are introduced. In Section 3.3 we initially focus on a simplified estimation problem, namely when the state sequence is measured directly, and apply the familiar least squares approach. Some convergence results are presented. When the state sequence is not measured directly, the least squares approach leads to the proposal of an ELS algorithm. We then generalize the ELS algorithm by introducing a RSPE scheme and demonstrate convergence via ODE analysis. In Section 3.4 an algorithm for the simultaneous estimation of transition probabilities $A$ and state output mappings $C$ is presented. In Section 3.5 some simulation studies that show relative performance of these algorithms are presented. Finally, conclusions are presented in Section 3.6.

### 3.2 Problem Formulation

In this section we introduce the HMM in state space form. Conditional state estimates and a parametrized information state model are also introduced.

#### 3.2.1 HMM State Space Model

Let $X_k$ be a discrete-time homogeneous, first-order Markov process, belonging to a finite set. The state space, $X$, *without loss of generality*, can be identified with a set of unit vectors, $S = \{e_1, e_2, \ldots, e_N\}$, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)' \in R^N$ with 1 in the $i$th position. We consider this process to be defined on the probability space $(\Omega, \mathcal{F}, P)$, with $\mathcal{F}_k^0 = \sigma\{X_0, \ldots, X_k\}$ and with complete filtration $\{\mathcal{F}_k\}$. The state space model is then defined, for $k \geq 0$, by

\[
X_{k+1} = AX_k + M_{k+1} \quad (3.2.1)
\]

\[
y_k = CX_k + w_k \quad (3.2.2)
\]

where $M_{k+1}$ is a sequence of $\mathcal{F}_k$-martingales and hence $E[M_{k+1}|\mathcal{F}_k] = 0$. Also the $y_k$ are continuous valued belonging to $R$ (although generalization to $R^N$ is straightforward) and $w_k \in R$ is *i.i.d.* with zero mean and of known density, such as when $w_k$ is Gaussian, i.e.
$w_k \sim N[0, \sigma_w^2]$, or a mixture of Gaussians. Also, $C \in \mathbb{R}^{1 \times N}$ is a vector of state values termed the state output mappings of the Markov chain. The term state values is commonly used for the vector $C$ when the observations are scalar. We also define $Y_k := (y_0, \ldots, y_k)$ and $\mathcal{Y}_k$ as the complete filtration generated by $y_\ell, \ell \leq k$. As a consequence,

$$E[w_{k+1} | \mathcal{F}_k \vee \mathcal{Y}_k] = 0. \quad (3.2.3)$$

Due to the Markov nature of $X_k$, we can write

$$E[X_{k+1} | \mathcal{F}_k] = E[X_{k+1} | X_k] = AX_k$$

where $A = (A^{ij})$ and $A^{ij} := P(X_{k+1} = e_i | X_k = e_j)$. Obviously, $A^{ij} \geq 0$ and $\sum_{i=1}^{N} A^{ij} = 1$, for all $j$. We also assume that $X_0$ or its distribution is known.

We shall define the vector of parametrized probability densities (or symbol probabilities) as $b_k = (b_k(i))$, for $b_k(i) := P[y_k | X_k = e_i]$. In the special case when $w_k \sim N[0, \sigma_w^2]$, we can write,

$$b_k(i) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp \left( -\frac{(y_k - Ce_i)^2}{2\sigma_w^2} \right). \quad (3.2.4)$$

We also write the initial state probability vector for the Markov chain $\pi = (\pi_i)$ with $\pi_i := P(X_0 = e_i)$. The HMM is denoted $\lambda = (A, C, \pi, \sigma_w^2)$.

### 3.2.2 Conditional State Estimates and Information State Model

Let $\hat{X}_{k|k,A}$ denote the conditional filtered state estimate of $X_k$, given measurements $Y_k$ and $A$. Also, let $\hat{X}_{k|k-1,A}$ denote the one-step-ahead prediction of $X_k$, given measurements $Y_{k-1}$ and $A$. That is,

$$\hat{X}_{k|k,A} := E[X_k | \mathcal{Y}_k, A], \quad \hat{X}_{k|k-1,A} := E[X_k | \mathcal{Y}_{k-1}, A]. \quad (3.2.5)$$

The forward recursion for obtaining conditional filtered state estimates $\hat{X}_{k|k,A}$ for an HMM is given in [18],

$$\hat{X}_{k|k,A} = N_k(y_k, A)B(y_k)A \hat{X}_{k-1|k-1,A}, \quad (3.2.6)$$

where $N_k(y_k, A) = \langle B(y_k)A \hat{X}_{k-1|k-1,A}, 1 \rangle^{-1}$ is a scalar normalization factor and $B(y_k) = \text{diag}([b_k(1), \ldots, b_k(N)])$ where $\text{diag}(X)$ is the diagonal matrix with $X$ on its diagonal.

We now proceed to introduce an information state model: An information state tells us all the information we know about the state from the observations, and is here simply the state estimate $\hat{X}_{k|k,A}$. Consider the following lemmas.
Lemma 3.2.1 The one-step-ahead predictions $E[X_k|Y_{k-1}, A]$ are given by

$$\tilde{X}_{k|k-1, A} = A\tilde{X}_{k-1|k-1, A}. $$

Proof

$$\tilde{X}_{k|k-1, A} = E[X_k|Y_{k-1}, A] = E[AX_{k-1} + M_k|Y_{k-1}, A] = A\tilde{X}_{k-1|k-1, A}. $$

Hence the following lemma now holds.

Lemma 3.2.2 The error term $(\tilde{X}_{k+1|k+1, A} - A\tilde{X}_{k|k, A})$ is orthogonal to $\tilde{X}_{k|k, A}$ and the error term $(y_k - C\tilde{X}_{k|k, A})$ is orthogonal to $\tilde{X}_{k|k, A}$. Moreover, the HMM can be reorganized as an information state model, see [41, page 79]. The state estimates can be written as

$$\tilde{X}_{k+1|k+1, A} = A\tilde{X}_{k|k, A} + n_k $$

$$y_k = C\tilde{X}_{k|k, A} + v_k $$

where $n_k$ and $v_k$ are orthogonal to $\tilde{X}_{k|k, A}$.

Proof From (3.2.1) we see that

$$\tilde{X}_{k+1|k+1, A} - A\tilde{X}_{k|k, A} = A[X_k - \tilde{X}_{k|k, A}] - [X_{k+1} - \tilde{X}_{k+1|k+1, A}] + M_{k+1}. $$

We next show that each term on the right is orthogonal to $\tilde{X}_{k|k, A}$.

From optimality, the estimation error $(X_k - \tilde{X}_{k|k, A})$ is orthogonal to $\tilde{Y}_k$ and since $\tilde{X}_{k|k, A} \subset \tilde{Y}_k$ then $\tilde{X}_{k|k, A}$ is orthogonal to $(X_k - \tilde{X}_{k|k, A})$. Similarly, $(X_{k+1} - \tilde{X}_{k+1|k+1, A})$ is orthogonal to $\tilde{Y}_{k+1}$ and because $\tilde{X}_{k|k, A} \subset \tilde{Y}_k \subset \tilde{Y}_{k+1}$ then $\tilde{X}_{k|k, A}$ is orthogonal to $(X_{k+1} - \tilde{X}_{k+1|k+1, A})$. Finally, $\tilde{X}_{k|k, A}$ is orthogonal to $M_{k+1}$ from (3.2.2) and because $\tilde{X}_{k|k, A}$ is orthogonal to $\tilde{Y}_k$. The result (3.2.7) follows, and (3.2.8) likewise by noting that $(y_k - C\tilde{X}_{k|k, A}) = C[X_k - \tilde{X}_{k|k, A}] + w_k$.

Lemma 3.2.2 shows that the orthogonality property required for convergence of standard recursive identification is satisfied, see [51].

3.3 Estimation of Transition Probabilities

In this section we develop algorithms for estimating the HMM transition probability matrix $A$ from observations $Y_k$. Initially, we investigate the simplified problem of estimating $A$ from a
known state sequence \(\{X_k\}\) using a least squares (LS) algorithm. In the following subsection, we use conditional state estimates in an extended least squares (ELS) algorithm to produce estimates of \(A\) when the state sequence is not measured directly. Finally, we introduce a state prediction error cost and propose a recursive (state) prediction error (RSPE) algorithm.

### 3.3.1 Least Squares

In this subsection, we consider the signal model (3.2.1),(3.2.2) and the simplified estimation problem: Estimate the state transition probability matrix \(A\) from the state sequence \(X_1, X_2, \ldots, X_k\). Subsequently, we will consider the more difficult estimation problem where the state sequence \(X_1, X_2, \ldots, X_k\) must be estimated from \(Y_k\).

**Lemma 3.3.1** Once each state has been active at least once, that is \((\sum_{k=1}^{m} X_k X_k')^{-1}\) exists, the optimal off-line least squares estimate of the transition probability matrix \(A\), given \(X_1, X_2, \ldots, X_m\), is

$$\hat{A}_m := \left( \sum_{k=1}^{m} X_{k+1} X_k' \right) \left( \sum_{k=1}^{m} X_k X_k' \right)^{-1}.$$  \hspace{1cm} (3.3.1)

Moreover,

$$\lim_{m \to \infty} \hat{A}_m \text{ exists a.s.}$$

Furthermore, under the excitation condition assumption \(\lim_{m \to \infty} (\sum_{i=1}^{m} X_i X_i')^{-1} = 0\), then

$$\lim_{m \to \infty} \hat{A}_m = A \quad \text{a.s..} \hspace{1cm} (3.3.2)$$

**Proof** Standard least squares algorithms are concerned with minimization with respect to \(A\) of the following cost,

$$\sum_{k=1}^{m} \| X_{k+1} - AX_k \|^2.$$ \hspace{1cm} (3.3.3)

Standard manipulations give (3.3.1). Now, since \((\sum_{k=1}^{m} X_k X_k') = (\sum_{k=1}^{m} \text{diag}(X_k))\), where \(\text{diag}(X)\) is the diagonal matrix with \(X\) on its diagonal when \(X\) is a vector, then

$$\hat{A}_{m}^{(i)} = \sum_{k=1}^{m} X_{k+1}^{(i)} X_k \left[ \sum_{k=1}^{m} X_k^{(j)} \right]^{-1}.$$\hspace{1cm} (3.3.4)

Also, since \(X_k^{(j)} \in \{0, 1\}\) then on the subsequence of \([1,m]\) for which \(X_k^{(j)} = 1\), denoted \(\{\ell_j(1), \ell_j(2), \ldots, \ell_j(m_j)\}\) with \(m_j\) integers (where \(m_j := \sum_{k=1}^{m} X_k^{(j)}\)), then

$$\hat{A}_{m}^{(i)} = m_j^{-1} \sum_{k=1}^{m_j} X_{\ell_j(k+1)} \leq 1, \text{ where } m_j \geq 1, \sum_{j=1}^{N} m_j = m.$$


First we prove the second lemma result (3.3.2), where the excitation condition, that
\[ \lim_{m \to \infty} \tilde{m}_j^{-1} = 0 \quad \text{for all } j, \]
holds. Consider the error term, which follows from algebraic manipulation of (3.3.1) and (3.2.1),
\[
\tilde{A}_m^j - A^j = \tilde{m}_j^{-1} \left[ \sum_{k=1}^{m_j} M_{(i)}^{(j)} \right].
\]

Now we define \( W_{m}^{(i,j)} := \sum_{k=1}^{m_j} \frac{1}{k} M_{(i)}^{(j)} \) whose elements are scalar martingales adapted to \( \mathcal{F}_{k} \) since \( E[W_{m+1}^{(i,j)} | \mathcal{F}_{m}] = W_{m}^{(i,j)} \) for all \( i, j \). Also, \( W_{m}^{(i,j)} \) is bounded in \( L_2 \) for each \( i, j \) since
\[
E \left[ (W_{m}^{(i,j)})^2 \right] &= E \left[ \left( \sum_{n=1}^{m_j} \frac{1}{n} M_{(i)}^{(j)} \right) \left( \sum_{k=1}^{m_j} \frac{1}{k} M_{(i)}^{(j)} \right) \right] \\
&= E \left[ \sum_{k=1}^{m_j} \frac{1}{k^2} E \left[ (M_{(i)}^{(j)}(k+1))^2 \right] | \mathcal{F}_{k} \right] \\
&\leq B_{\infty} \sum_{k=1}^{m_j} \frac{1}{k^2} < \infty \quad \text{for all } i, j.
\]

Here we have used that \( E(M_{k+1}^{(i)} M_{n}^{(i)} | \mathcal{F}_{k}) = 0 \) for all \( n \leq k \) and \( E((M_{k+1}^{(i)})^2 | \mathcal{F}_{k}) \leq B_{\infty} \) for some \( B_{\infty} < \infty \).

Now under the excitation condition \( \tilde{m}_j \to \infty \) for all \( j \), and martingale convergence results [59, 64] we have that \( W_{m}^{(i,j)} \) converges almost surely. Hence by the Kronecker Lemma [64, 53] we have that
\[
\lim_{m \to \infty} (\tilde{A}_m^j - A^j) = \lim_{m_j \to \infty} \frac{1}{\tilde{m}_j} \sum_{k=1}^{m_j} M_{(i)}^{(j)}(k+1) = 0 \quad \text{a.s. for all } i, j
\]
and the lemma result (3.3.2) follows.

To obtain the first lemma result we note that if \( \lim_{m \to \infty} \tilde{m}_j \) is finite then \( \sum_{k=1}^{m_j} X_{(i)}^{(j)}(k+1) \) is also finite, and hence clearly \( \lim_{m \to \infty} \tilde{A}_m \) is finite. Existence of \( \lim_{m \to \infty} \tilde{A}_m \) when \( \lim_{m \to \infty} \tilde{m}_j^{-1} = 0 \) is proven by the second lemma result hence the first lemma result follows as claimed.

Consider now on-line estimation via recursive least squares (RLS) algorithms. Simple manipulations of (3.3.1) give the on-line recursions
\[
\tilde{A}_{k+1} = \tilde{A}_k + (X_{k+1} - \tilde{A}_k X_k) X_k^P k
\]
\[ P_k^{-1} = P_{k-1}^{-1} + X_k X_k', \quad \text{or} \]
\[ P_k = P_{k-1} - P_{k-1} X_k [1 + X'_k P_{k-1} X_k]^{-1} X'_k P_{k-1} \]  
(3.3.4)

where \( P_k \) can be thought of as related to the energy of the input sequence.

The indicator vectors \( X_k \) have the property that non-linear functions of an indicator vector \( F(X_k) \) are linear functions \([F(e_1), \ldots, F(e_N)]X_k \) of the indicator vector \( X_k \). Exploiting this property it is possible to rewrite (3.3.4) so that the right hand sides are linear in \( X_k \).

We now proceed to consider the more realistic case when \( X_k \) is not measured directly, but must be estimated from observations. We first examine extended least squares (ELS) algorithms.

### 3.3.2 Extended Least Squares

This subsection proposes an ELS algorithm for estimating HMM transition probabilities. Extended least squares algorithms are ad hoc algorithms in which conditional state estimates are used in lieu of actual states \( X_k \) in an LS implementation, see [72] for more details.

Consider the ELS version of the LS recursion (3.3.4) obtained by replacing the state \( X_k \) by conditional state estimates, that is,

\[
\begin{align*}
\widehat{A}_{k+1} &= \widehat{A}_k + (\widehat{X}_{k+1|k+1}, \widehat{A}_u - \widehat{A}_k \widehat{X}_{k|k}, \widehat{A}_{k-1}) \widehat{X}'_{k|k}, \widehat{A}_{k-1} \widehat{P}_k^{-1} \\
\widehat{P}_k^{-1} &= \widehat{P}_{k-1}^{-1} + \text{diag}(\widehat{X}_{k|k}, \widehat{A}_{k-1}) 
\end{align*}
\]  
(3.3.5)

where \( \widehat{A}_k = [\widehat{A}_0, \ldots, \widehat{A}_k] \), \( \text{diag}(X) \) is the diagonal matrix with \( X \) on its diagonal when \( X \) is a vector, and the recursion below is used to generate state estimates, \( \widehat{X}_{k+1|k+1}, \widehat{A}_{k-1} \)

\[ \widehat{X}_{k+1|k+1}, \widehat{A}_{k} = N_{k+1}(y_k, \widehat{A}_k) B(y_{k+1}) \widehat{A}_k \widehat{X}_{k|k}, \widehat{A}_{k-1} \]  
(3.3.6)

where \( N_{k+1}(y_k, \widehat{A}_k) \) is a scalar normalization factor as in (3.2.6).

**Remarks**

1. Note that \((\widehat{X}_{k+1|k+1}, \widehat{A}_u - \widehat{A}_k \widehat{X}_{k|k}, \widehat{A}_{k-1})\) is not orthogonal to \( \widehat{X}_{k|k}, \widehat{A}_{k-1} \) unless \( \widehat{A}_k = A \) for all \( k \). Hence, standard theory no longer applies.

2. The computational cost of the ELS recursion (3.3.5) at each iteration is \( O(N^2) \).
Since we are unable to proceed with further analysis of the convergence properties of this ELS algorithm we proceed in the next subsection by taking the ELS concepts one step further. The RSPE (recursive state prediction error) algorithm which follows appears to naturally generalize this ELS algorithm. These RSPE algorithms are developed with the view of achieving asymptotic efficient convergence (in the sense of almost surely to a local minimum of the appropriate cost function) with rate of order $1/k^4$.

### 3.3.3 Recursive State Prediction Error Method

There exists mature theory for recursive identification of discrete-time models with states in $R^N_+$ based on the minimization of the observation prediction error cost, see [51]. This RPE theory provides asymptotic quadratic convergent algorithms (admittedly to a local minimum) for linear and certain non-linear models.

In this section we proceed by applying this theory to obtain asymptotic convergent algorithms (in a local sense) for HMM identification which generalize the ELS scheme of the previous subsection.

Lemma 3.2.2 motivates the use of a state prediction error cost (see (3.3.3)), rather than the observation prediction error cost that is used in the standard RPE theory. Consider the cost function

$$V_k(\theta) := \frac{1}{2} \sum_{i=2}^{k} \| \tilde{X}_{i|\theta} - A(\theta)\tilde{X}_{i-1|\theta} \|^2,$$  \hspace{1cm} (3.3.7)

where $\theta$ is used to parametrize the unknown transition probability matrix such that, $\theta = [\mathbf{a}_1, \ldots, \mathbf{a}_N]'$ where $\mathbf{a}_i := [A_{i1}, \ldots, A_{iN}]$.

Thus the RSPE recursions which seek to minimize the cost (3.3.7) are

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \kappa_k|\hat{\theta}_{k-1}$$

$$\hat{P}_{k}^{-1} = \hat{P}_{k-1}^{-1} + \text{diag}(1' \otimes \tilde{X}_{k-1|\theta_{k-1}}, \tilde{X}_{k-1|\theta_{k-1}}, \tilde{X}_{k-1|\theta_{k-1}}), \quad \hat{P}_0^{-1} = \Delta I$$  \hspace{1cm} (3.3.8)

where $\kappa_k|\hat{\theta}_{k-1} = \kappa_k|\hat{\theta}_{k-1}$, for $\kappa_k|\hat{\theta}_{k-1} = \frac{\partial}{\partial \theta_k} V_k(\theta) |_{\theta = \hat{\theta}_{k-1}}$, $1'$ is defined as the column vector of all ones, $\otimes$ is the Kronecker product, and $\Delta$ is some large constant. Here $\hat{P}_{k-1}^{-1}$ is an approximation to the second derivative of $V_k(\theta)$. Note that a projection operation can be implemented at each time step to ensure that $A(\hat{\theta}_k)$ is a valid stochastic matrix and the convergence results presented in the following discussion still hold.
The recursion (3.3.8) can also be written as the following scalar recursion

\[
\tilde{\theta}_k^{(i)} = \tilde{\theta}_{k-1}^{(i)} + \tilde{P}_k^{(i)} \kappa_k^{(i)} \tilde{\theta}_{k-1}^{(i)} \\
(\tilde{P}_k^{(i)})^{-1} = (\tilde{P}_{k-1}^{(i)})^{-1} + \tilde{X}_{k-1}^{(i)} \tilde{X}_{k-1}^{(i)\top} \tag{3.3.9}
\]

where \( \xi = i \mod n \). Here, \( \mod \) is the usual modulo operation except that \( i \mod n = i \).

**Gradient Calculations**

\[
\kappa_k^{(i)} \tilde{\theta}_{k-1}^{(i)} = \frac{dV_k(\theta)}{dA^{mn}} \bigg|_{\theta = \tilde{\theta}_{k-1}}
\]

where \( m = i \mod n \) and \( n = (i - m)/n + 1 \). Now

\[
\frac{dV_k(\theta)}{dA^{mn}} = - \left( \tilde{X}_{k|k,\theta}^{(n)} - a_m(\theta)\tilde{X}_{k-1|k-1,\theta}^{(n)} \right) \tilde{X}_{k-1|k-1,\theta}^{(n)\top} + \\
\sum_{j=1}^{N} \left( \tilde{X}_{k|k,\theta}^{(j)} - a_j(\theta)\tilde{X}_{k-1|k-1,\theta}^{(j)} \right) \times \\
\left( \frac{d\tilde{X}_{k|k,\theta}^{(j)}}{dA^{mn}} - a_j(\theta) \frac{d\tilde{X}_{k-1|k-1,\theta}^{(j)}}{dA^{mn}} \right). \tag{3.3.10}
\]

Here \( a_j \) is defined in section 3.2.1, and

\[
\frac{d\tilde{X}_{k|k,\theta}^{(j)}}{dA^{mn}} = N_k^2 \text{diag}(B) \left( \tilde{X}_{k-1|k-1,\theta}^{(n)} \sigma_m + A(\theta) \frac{d\tilde{X}_{k-1|k-1,\theta}^{(n)}}{dA^{mn}} \right) \times \\
B(j, j) a_j(\theta) \tilde{X}_{k-1|k-1,\theta}^{(j)} + \\
N_k B(j, j) a_j(\theta) \frac{d\tilde{X}_{k-1|k-1,\theta}^{(j)}}{dA^{mn}} \text{ if } j \neq m
\]

\[
\frac{d\tilde{X}_{k|k,\theta}^{(j)}}{dA^{mn}} = N_k^2 \text{diag}(B) \left( \tilde{X}_{k-1|k-1,\theta}^{(n)} \sigma_m + A(\theta) \frac{d\tilde{X}_{k-1|k-1,\theta}^{(n)}}{dA^{mn}} \right) \times \\
B(j, j) a_j(\theta) \tilde{X}_{k-1|k-1,\theta}^{(j)} + \\
N_k B(j, j) \left( \tilde{X}_{k-1|k-1,\theta}^{(n)} + a_j(\theta) \frac{d\tilde{X}_{k-1|k-1,\theta}^{(n)}}{dA^{mn}} \right) \text{ if } j = m. \tag{3.3.11}
\]

**Convergence Proof**

Convergence of (3.3.8) and (3.3.9) is shown by considering the ordinary differential equation (ODE) associated with (3.3.8) and (3.3.9). That is,

\[
\frac{d}{d\tau} \theta(\tau, k) = R^{-1}(\tau, k)f(\theta(\tau, k), k) \\
\frac{d}{d\tau} R(\tau, k) = G(\theta(\tau, k), k), \quad R(0, k) \geq \delta I. \tag{3.3.12}
\]
Here $k$ is fixed and $\delta$ is a small constant. Let us define for (3.3.8) and (3.3.9), with $\theta(\tau, k)$ abbreviated as $\theta_{\tau}$,

$$f(\theta_{\tau}, k) = E[k_{k|\theta_{\tau}}] \quad \text{and}$$

$$G(\theta_{\tau}, k) = E[\text{diag}(\mathbf{1}' \otimes \hat{X}_{k-1|\theta_{\tau}})].$$

(3.3.13) (3.3.14)

The following lemma now holds.

**Lemma 3.3.2** The recursions (3.3.8) and (3.3.9) will converge a.s. to the set

$$\tilde{D} = \{ \theta_{\infty} | \lim_{k \to \infty} E[f(\theta_{\infty}, k)] = 0 \} \supset \theta \quad \text{(or possibly the boundary of the valid A region if a projection step is performed).}$$

Moreover, under the excitation condition $\tilde{P}_k \to 0$ as $1/k$, then convergence of $\tilde{\theta}_k$ is at the rate $1/k^{1/2}$.

**Proof** The ODE associated with (3.3.8) and (3.3.9) for fixed $k$, under (3.3.13) and (3.3.14), is (3.3.12).

Now a Lyapunov function for (3.3.12) under (3.3.13) and (3.3.14), is

$$\tilde{W}(\theta_{\tau}, k) = E[||\hat{X}_{k|\theta_{\tau}} - A(\theta_{\tau})\hat{X}_{k-1|\theta_{\tau}}||^2]$$

(3.3.15)

so that

$$\frac{d}{d\tau} \tilde{W}(\theta_{\tau}, k) = \frac{d\tilde{W}(\theta_{\tau}, k)}{d\theta_{\tau}} \frac{d\theta_{\tau}}{d\tau}$$

$$= -f'(\theta(\tau, k), k)R(\tau, k)^{-1}f(\theta(\tau, k), k).$$

(3.3.16)

Thus $\tilde{W}(\theta_{\tau}, k)$ converges for all $k$ and $\tau \to \infty$, and $\theta(\tau, k)$ converges to the set $\{ \theta | E[f(\theta, k)] = 0 \}$ (for discussion of convergence when a projection is performed, see Ljung [51]).

Here the recursions (3.3.8) and (3.3.9) and intermediate steps are stable, hence together with the results of [8, 70, 45] the various regularity conditions required by the ODE theory of Ljung [51] are satisfied and the first result claimed follows. Note the conditions given in [8, 70, 45] ensure that HMM filters forget initial conditions exponentially.

Observe from (3.3.16) that if $R(\tau, k)$ is of the order $1/k$, as under suitable excitation, then $f(\theta(\tau, k), k)$ converges to zero as $1/k^{1/2}$. Since asymptotically the stochastic difference equation behaves as the ODE, then rates of convergence translate across.

This leads to the convergence rate result of the lemma.

□
§3.3 Estimation of Transition Probabilities

Remarks

1. The theory is not a global convergence theory. It is not excluded that the set $\mathcal{D}$ may contain local optima. Simulation studies suggest that with reasonable initializations, $\hat{\theta}_k$ converges to $\theta$, as desired.

2. The lemma excitation condition $\hat{P}_k \to 0$ as $1/k$ is not particularly restrictive. It can be interpreted as an ergodicity requirement on the state sequence. That is, the Markov state sequence must visit each state (uniformly) as $k \to \infty$.

3. The existence of parameter estimates and/or convergence of these estimates (possibly only for a subset of the parameters) can be shown when the lemma excitation condition is relaxed, but this is not done here.

4. To reduce the number of calculations the second half of (3.3.8) and (3.3.9) can be replaced by a stochastic approximation given by

$$\hat{P}_k^{-1} \sim k \text{diag}(E[X_k]).$$

Convergence can still be proven with a slight modification of Lemma 3.3.2.

5. The concept of using a cost function (3.3.7) that measures the state prediction error has been introduced previously in other contexts by Bryson, see [7, Page 349]. However, we believe this concept has not been used previously for HMM identification.

6. The state prediction error can not be driven to zero for all $k$ by a particular choice of $\theta$ due to the nature of Markov sequences, however, the expected value of the error will tend to zero as $\hat{\theta}_k \to \theta$.

7. The number of calculations required to estimate $\theta$ in (3.3.8) and (3.3.9) is of $O(N^4)$.

In [13], the observation prediction cost function is used to identify transition probabilities, that is,

$$\hat{\theta} = \arg \min_{\theta} \{ \hat{V}_k(\theta) = E [(y_k - CX_k)^2|Y_{k-1}] \}.$$

To understand the difficulty in using this type of cost function to estimate the transition probabilities of an HMM consider the following lemma.
Lemma 3.3.3 As the measurement noise approaches zero in variance, that is \( \sigma_w^2 \to 0 \), then
\[
\frac{d\hat{X}^{(j)}_{k|k-1,\tilde{\omega}_{k-1}}}{d\mathbf{A}^m} \to 0.
\]

Proof From (3.2.6) we see that,
\[
\hat{X}^{(j)}_{k+1|k+1,\tilde{\omega}_k} = N_k b_k(j) a_j \hat{X}^{(j)}_{k|k,\tilde{\omega}_{k-1}}
\]
where \( N_k = \left( \sum_{j=1}^N b_k(j) a_j \hat{X}^{(j)}_{k|k,\tilde{\omega}_{k-1}} \right)^{-1} \) and \( b_k(i) \) is defined in (3.2.4).

As \( \sigma_w^2 \to 0 \) then \( b_k(i) \to 0 \) for all \( i \) that \( X_k \neq e_i \) and \( b_k(i) \neq 0 \) for the \( i \) that \( X_k = e_i \).

Hence, \( N_k \to b_k(i) a_i \hat{X}^{(i)}_{k|k,\tilde{\omega}_{k-1}} \) a.s. for the \( i \) that \( X_k = e_i \).

Thence, \( \hat{X}^{(i)}_{k+1|k+1,\tilde{\omega}_k} \to 0 \) for all \( X_k \neq e_i \) and \( \hat{X}^{(i)}_{k+1|k+1,\tilde{\omega}_k} \to 1 \) for the \( i \) that \( X_k = e_i \), i.e. \( \hat{X}^{(i)}_{k+1|k+1,\tilde{\omega}_k} \to X_{k+1} \) The lemma result follows.

Lemma 3.3.3 implies that \( \frac{d}{d\theta} \tilde{V}_k(\theta) \to 0 \) as \( \sigma_w^2 \to 0 \). That is, as \( \sigma_w^2 \to 0 \) the cost function \( \tilde{V}_k(\theta) \) becomes invariant of \( \theta \). Hence, it is clear that \( \tilde{V}_k(\theta) \) is not a good criterion for identifying \( \theta \). Lemma 3.3.3 correctly predicts that the performance of the RPE algorithm presented in [13] will deteriorate as \( \sigma_w^2 \to 0 \).

Our choice of cost function (3.3.7) does not suffer from the same difficulties as \( \sigma_w^2 \to 0 \). In fact, from (3.3.10) it is clear that as \( \sigma_w^2 \to 0 \) the RSPE algorithm reduces to the ELS algorithm (3.3.5). Similarly, as \( \sigma_w^2 \to 0 \) then \( \hat{X}^{(i)}_{k+1|k+1,\tilde{\omega}_k} \to X_k \) and hence the ELS algorithm, and likewise the RSPE algorithm, simplifies to the LS algorithm (3.3.4).

Remark

1. Even without \( \sigma_w^2 \to 0 \) it is possible to see the similarities between the ELS recursion (3.3.5) and the RSPE recursion (3.3.8). In fact, if we were to approximate the gradient \( \kappa_{k|\tilde{\omega}_{k-1}} \) by the first term in (3.3.10) then the RSPE recursions would reduce to the ELS recursions (3.3.5).

3.4 Simultaneous Estimation

This section proposes an algorithm for simultaneous estimation of the state output mapping matrix \( C \) and the transition probability matrix \( A \), given a set of observations \( Y_k \) and knowledge
of the measurement noise variance $\sigma_m^2$. Local convergence results are presented. Stronger convergence results are not shown nor excluded from our theory.

### 3.4.1 Dual Cost Function Approach

To obtain simultaneous estimates for $A$ and $C$ we consider the coupled sub-problems of estimating $C$ given an estimate of $A$, and estimating $A$ given an estimate of $C$. Each of these sub-problems can be solved respectively via RPE and RSPE techniques after setting up appropriate cost functions. The estimates from the C-recursion and A-recursion can be fed back into the A-recursion and C-recursion respectively to couple the recursions.

Consider the minimization of the two separate cost functions, (3.4.1) and (3.4.2).

$$\tilde{\theta}_k^C = \operatorname*{arg\,min}_{\theta^C} \left\{ V_1^C(\theta^C, \tilde{\theta}_k^{A-1}) = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{N} (y_i - C(\theta^C)e_j)^2 P(X_i = e_j | y_i) \right\} \tag{3.4.1}$$

$$\tilde{\theta}_k^A = \operatorname*{arg\,min}_{\theta^A} \left\{ V_2^A(\theta^A, \tilde{\theta}_k^{C-1}) = \frac{1}{2} \sum_{i=2}^{k} \| \hat{X}_{i|\hat{\theta}_{i-1}} - A(\theta^A) \hat{X}_{i-1|\hat{\theta}_{i-2}} \|^2 \right\}. \tag{3.4.2}$$

Here the two parametrations $\theta^A := [a_1, \ldots, a_N]'$ and $\theta^C := [C^{(1)}, \ldots, C^{(N)}]'$ have been introduced and $\hat{\Theta}_k := [\hat{\theta}_0^A, \hat{\theta}_0^C, \ldots, \hat{\theta}_k^A, \hat{\theta}_k^C]$ denotes the history of estimation. The cost functions $V_1^C(\theta^C, \tilde{\theta}_k^{A-1})$ and $V_2^A(\theta^A, \tilde{\theta}_k^{C-1})$ are coupled through the $\tilde{\theta}_k^A$ and $\tilde{\theta}_k^C$ terms.

We proceed by introducing recursions in $\tilde{\theta}_k^C$ and $\tilde{\theta}_k^A$ before establishing convergence results.

$$\tilde{\theta}_k^C = \tilde{\theta}_k^{C-1} + \hat{P}_k \phi_k |_{\tilde{\theta}_k^{C-1}} \tilde{\theta}_k^{A-1}$$

$$\tilde{\theta}_k^{-1} = \tilde{\theta}_k^{-1} + \operatorname{diag}(P(X_k = e_1 | y_k), \ldots, P(X_k = e_N | y_k)) \tag{3.4.3}$$

where $\phi_k |_{\tilde{\theta}_k^{C-1}} = \frac{\partial}{\partial \hat{\theta}_k^C} V_1^C(\theta^C, \tilde{\theta}_k^{A-1}) |_{\hat{\theta}_k^C = \tilde{\theta}_k^{C-1}}$, and

$$\tilde{\theta}_k^A = \tilde{\theta}_k^{A-1} + \hat{P}_k \kappa_k |_{\tilde{\theta}_k^{A-1}} \tilde{\theta}_k^{C-1}$$

$$\tilde{\theta}_k^{-1} = \tilde{\theta}_k^{-1} + \operatorname{diag}(I' \otimes \hat{X}_{k-1|\hat{\theta}_{k-1}, \hat{\theta}_{k-2}}) \tag{3.4.4}$$

where $\kappa_k |_{\tilde{\theta}_k^{A-1}}$ is the same as $\kappa_k |_{\tilde{\theta}_k^{C-1}}$ defined in (3.3.10)

### Convergence Proof

To demonstrate local convergence of the coupled algorithm we first show that that recursion (3.4.3) converges locally independently of $\tilde{\theta}_k^A$ (or recursion (3.4.4)). Next we show local convergence of recursion (3.4.4).
Lemma 3.4.1 If the parametrized probability densities \( b_k \) are independent of \( \theta^A \) then the cost function (3.4.1) is independent of \( \tilde{\theta}^A_i \), \( i = 0, \ldots, k \).

Proof The lemma condition implies that \( P(X_i = e_j|Y_i) \) (as distinct from \( P(X_i = e_j|Y_i) \)) is independent of \( \tilde{\theta}^A_{k-1} \), hence the cost \( V_k^1(\theta^C, \tilde{\theta}^A_{k-1}) \) in (3.4.1) is independent of \( \tilde{\theta}^A_{k-1} \). □

It follows from Lemma 3.4.1 that the recursions (3.4.3) are independent of \( \tilde{\theta}^A_{k-1} \) and hence convergence of (3.4.3) can be established as follows.

Consider the ODE (3.3.12) and with \( \theta(\tau, k) \) abbreviated as \( \theta^C \) and let us redefine for (3.4.3) the following,

\[
\begin{align*}
f(\theta^C, k) &= E[\phi_k|\theta^C] \quad \text{and} \\
G(\theta^C, k) &= E[\text{diag}(\tilde{X}_{k-1}|\theta^C, \tilde{\theta}^A_{k-1})].
\end{align*}
\]

The following lemma holds.

Lemma 3.4.2 If the parametrized probability densities \( b_k \) are independent of \( \theta^A \), then the recursion (3.4.3) will converge a.s. to the set \( \tilde{D}_{\theta^C} = \{\theta^C|\lim_{\tau \to \infty} E[f(\theta^C, k)] = 0\} \supset \theta^C \). Moreover, under the excitation condition \( \hat{P}_k \to 0 \) as \( 1/k \), then convergence of \( \tilde{\theta}^C_k \) is at the rate \( 1/k^{1/2} \).

Proof A similar approach to Lemma 3.3.2 can be taken. See also [13]. □

Lemma 3.4.2 demonstrates local convergence results for the recursion (3.4.3). We now present convergence results for (3.4.4) under the assumption that (3.4.3) converges to the true value of \( \theta^A \). Again, consider the ODE (3.3.12) and with \( \theta(\tau, k) \) now abbreviated as \( \theta^A \) let us redefine for (3.4.4)

\[
\begin{align*}
f(\theta^A, k) &= E[k_k|\theta^A] \quad \text{and} \\
G(\theta^A, k) &= E[\text{diag}(\tilde{X}_{k-1}|\theta^A, \theta^A_{k-1})].
\end{align*}
\]

The following lemma now holds.

Lemma 3.4.3 Given that \( \tilde{\theta}^{C_k} \to \theta^C \) converges a.s. as \( k \to \infty \), then the recursion (3.4.4) will converge a.s. to the set \( \tilde{D}_A = \{\theta^A|\lim_{\tau \to \infty} E[f(\theta^A, k)] = 0\} \supset \theta^A \) (or possibly the boundary of the valid \( A \) region if a projection step is performed). Moreover, under the excitation condition \( \hat{P}_k \to 0 \) as \( 1/k \), then convergence of \( \tilde{\theta}^A_k \) is at the rate \( 1/k^{1/2} \).
Proof Because $\hat{\theta}_k^C \rightarrow \theta^C$ as $k \rightarrow \infty$, then likewise the cost $V^2_k(\theta^A, \hat{\theta}_k^C) \rightarrow V^2_k(\theta^A, \theta^C)$ as $k \rightarrow \infty$. Now, by inspection it is clear that $V^2_k(\theta^A, \theta^C)$ is equivalent to $V_k(\theta)$ given in Section 3.3. Hence, the rest of the proof follows Lemma 3.3.2 \hfill \Box

Together, Lemmas 3.4.1, 3.4.2 and 3.4.3 imply local convergence of parameter estimates $\hat{\theta}_k^C$ and $\hat{\theta}_k^A$. However, note that Lemma 3.4.3 holds if and only if (3.4.3) has converged to the true value of $\theta^C$ rather than locally as Lemma 3.4.2 provides. In particular, for noise processes which are multi-modal such as mixtures of Gaussian this may not always occur.

Remarks

1. Alternative cost function for estimating $C$ have been proposed elsewhere, see [13] and Chapter 2.

2. The Lemma 3.4.1 conditions are not very restrictive. For example, Gaussian noise models and mixtures of Gaussians noise models both satisfy the lemma condition.

3. $P(X_i = e_j|y_i)$ can be replaced by $P(X_i = e_j|Y_i)$ in the cost function (3.4.1), however, convergence is no longer guaranteed. In simulations it is found that a scheme with $P(X_i = e_j|Y_i)$ replacing $P(X_i = e_j|y_i)$ converges for all but the worst initial guesses. Note, that if $\hat{\theta}_k^A = [1/N^2, ..., 1/N^2]'$ then $P(X_i = e_j|y_i) = P(X_i = e_j|Y_i)$ making $\hat{\theta}_0^A = [1/N^2, ..., 1/N^2]'$ a good initialization for the modified scheme if no other a prior information is available.

4. The dual cost function approach of this section has been found in simulations to converge more rapidly than a composite single cost function approach, eg, minimization of $V_k(\theta) = V^1_k(\theta^C, \theta^A) + \lambda V^2_k(\theta^A, \theta^C)$, for some $\lambda$.

5. Implementation of recursions (3.4.3) and (3.4.4) requires $O(N^4) + O(N)$ calculations per time instant, which is less than the $O(N^4) + O(N^2)$ required using a composite single cost function approach. Further reduction in computational requirements can be achieved by implementing ELS versions of (3.4.3) and (3.4.4); however, convergence results are not yet established in this case.
3.5 Simulations

3.5.1 Implementation Considerations

In Chapter 2 several implementation issues are discussed, including:

- The use of step sequences and Polyak acceleration to improve transients performance.

- The modification of the parameter estimate recursions to include the variance of Markov state estimates and vice versa.

- Modifications to allow tracking of slowly time varying parameters.

The discussion in Chapter 2 equally applies to the algorithms presented in this chapter.

3.5.2 Simulation Results

We present results of simulation examples using computer generated finite, discrete-state Markov chains to demonstrate features of the algorithms proposed in this chapter.

Estimation of Transition Probabilities

A two-state Markov chain embedded in WGN is generated with parameter values \( A^{ii} = 0.9, A^{ij} = (1 - A^{ii}) \) for \( i \neq j \), \( C = [1, 3]' \), \( \sigma_w^2 = 1 \), assuming \( C \) and \( \sigma_w^2 \) known. The transition probability matrix is estimated using both the ELS and RSPE algorithms, (3.3.5) and (3.3.8), respectfully. Figure 3.1 shows a comparison of the estimation errors. This figure shows that convergence towards the true value occurs for both schemes and suggest that in this example the RSPE scheme converges more rapidly that the ELS scheme.

Estimation in low noise

A two-state Markov chain embedded in WGN is generated with parameter values \( A^{ii} = 0.9, A^{ij} = (1 - A^{ii}) \) for \( i \neq j \), \( C = [1, 3]' \), \( \sigma_w^2 = 0.0001 \), assuming \( C \) and \( \sigma_w^2 \) are known. The transition probabilities of the chain are estimated in low noise using the ELS algorithm, ie. (3.3.5). For this noise level, the recursive schemes presented in [13] do not converge. Figure 3.2 shows the error in estimation of (3.3.5) over time. This figure demonstrates that (3.3.5) converges occurs in this low noise example.
§3.6 Conclusions

Estimation of fast Markov chains

A two-state Markov chain embedded in WGN is generated with parameter values \( A^{ii} = 0.6, A^{ij} = (1 - A^{ii}) \) for \( i \neq j \), \( C = [1, 3]' \), \( \sigma_w^2 = 1 \), assuming \( C \) and \( \sigma_w^2 \) are known. The transition probabilities of the chain are estimated using the RSPE algorithm, ie (3.3.8). Figure 3.3 shows the size of the estimation error over time, and demonstrates that convergence occurs for this example.

Higher order chain

A three-state Markov chain embedded in WGN is generated with parameter values \( A^{ii} = 0.9, A^{ij} = (1 - A^{ii})/2 \) for \( i \neq j \), \( C = [1, 3, 5]' \), \( \sigma_w^2 = 1 \), assuming \( C \) and \( \sigma_w^2 \) are known. The transition probabilities of the chain are estimated using the RSPE algorithms, (3.3.8). Figure 3.4 shows the time evolution of the transition probabilities estimates. This figures demonstrates that estimates converge to the correct values.

Simultaneous Estimation

A two-state Markov chain embedded in WGN is generated with parameter values \( A^{ii} = 0.8, A^{ij} = (1 - A^{ii}) \) for \( i \neq j \), \( C = [1, 3]' \), \( \sigma_w^2 = 1 \) with \( \hat{\theta}_0^C = [0.5, 1] \) and \( \hat{\theta}_0^A = 0.5 \forall i \). The transition probabilities and the state output mappings of the chain are estimated simultaneously using (3.4.3) and (3.4.4). Figures 3.5 and 3.6 shows the time evolution of the transition probabilities and state values estimates respectfully. Figure 3.7 show the estimation error in \( C^{(1)} \) and the transition probability \( A^{11} \). These figures demonstrate that estimates converge to the correct values. Comparison with the results presented in [13] suggest that the convergence is considerably more rapid.

3.6 Conclusions

In this chapter we have proposed new algorithms for recursive estimation of the state transition probabilities for HMMs based on ELS and RSPE techniques. These algorithms avoid the ill-conditioning in low noise of the schemes in [13]. Convergence analysis for the RSPE algorithm is provided via an ODE approach but no convergence results are presented for the ELS algorithm. Despite the lack of convergence results the ELS algorithm is attractive because it has computational complexity of only \( O(N^2) \) per time instant, compared with the RPE
scheme (of [13]) and the RSPE scheme of this chapter which have computational complexity $O(N^4)$.

This chapter also proposes a scheme for the simultaneous estimation of state output mapping values and the state transition probabilities. Local convergence results are presented. The simulation studies presented illustrate the schemes proposed in this chapter and show examples where convergence from reasonable initializations and convergence in low noise levels occurs.
Figure 3.1: Comparison of convergence rates

Figure 3.2: Convergence in low noise
Figure 3.3: Convergence of a fast chain

Figure 3.4: Convergence of higher order chain
Figure 3.5: Simultaneous estimation: A estimates

Figure 3.6: Simultaneous estimation: C estimates
Figure 3.7: Simultaneous estimation: Estimation error
Chapter 4

Almost Sure Estimation of Hidden Markov Models

4.1 Introduction

The classical Baum-Welch algorithm works on finite data sets, see [4, 55]. It can be used to estimate the probability of a jump from state $i$ to state $j$, at each time $\ell$ for $0 \leq \ell \leq T$, using multiply forward and backward passes through the data set. The Baum-Welch re-estimation works with an estimated model which is updated after each forward-backward pass, and so is not really an on-line scheme. Also, the theory gives only convergence to a local maximum of the likelihood function.

In earlier results, adaptive HMM schemes have been proposed using on-line pseudo-likelihood maximization, which can therefore be considered as on-line versions of the Baum-Welch scheme, [39]. For these, the computational effort is of order $N^2$, and only local convergence results have been claimed.

Subsequently, in [13], adaptive HMM schemes based on recursive prediction error methods used commonly in the identification of linear systems have been proposed. These require $O(N^4)$ computational effort and have certain weaknesses in low noise, (surprisingly). The weaknesses of [13] have been avoided in the schemes proposed in chapters 2 and 3, by using a different prediction error cost. These latter schemes also exploit martingale convergence results to rationalize some of the previous schemes. These schemes are of order $(N^2)$.

In [55] it is shown that a standard technique for handling off-line least squares problems with missing data points is the Healy-Westmacott procedure, which can be viewed as an EM algorithm. Surprisingly, this connection also appears in the on-line case because the
on-line EM-like algorithm we present here and the recursive prediction error algorithms of
previous chapters appear connected through a Lyapunov function. Note that it is possible to
consider the recursive prediction error algorithms of previous chapters as on-line least squares
algorithms modified to handle a partially observed process.

In [18], we give recursive filters which yield conditional mean estimates of $\mathcal{J}_k^{ij}, \mathcal{T}_k^i$ given
the observations, where $\mathcal{J}_k^{ij} = \sum_{t=1}^k X_{l-1}^i X_t^j$ and $\mathcal{T}_k^i = \sum_{t=1}^k y_t X_t^i$. These can be exploited
for an on-line identification scheme. However, the filter update for each $\mathcal{J}_k^{ij}$ requires $N^2$
multiplications at each time step. There are $N^2$ different processes $\mathcal{J}_k^{ij}$, for $1 \leq i, j \leq N$;
consequently the update of the matrix $\mathcal{J}_k := (\mathcal{J}_k^{ij})$ requires $N^4$ multiplications at each time,
which is perhaps prohibitive for large $N$. There are $N^2$ multiplications required to update
the $\mathcal{T}_k^i$.

This chapter first addresses the issue of global convergence for a simplified estimation task
where full observation information is available. As a next step towards achieving global con-
vergence, results are presented for the idealized situation of using conditional mean estimates
of the states in lieu of the true states. This is an idealized situation since the estimates are
based on the true model which is actually unknown. It is shown for homogeneous Markov
models, via martingale convergence theorems, that the estimation schemes for the above sim-
plified estimation tasks tune to the optimal estimates almost surely at a guaranteed rate of
$(k^{-1} \ln(k \ln k)^{\rho})^{\frac{1}{\rho}}$, for any $\rho > 1$, where $k$ is the number of iterations. For mean square
convergence, the rate is arbitrarily slower than $k^{-\frac{1}{2}}$.

Global convergence results are then presented for an estimation algorithm that uses con-
ditional mean estimation based, not on the true model, but on an adaptive model estimate.
To show convergence the ordinary differential equation (ODE) approach of [50, 51, 42, 29] is
used. The global convergence results contrast the off-line Baum-Welch re-estimation results
which are only guaranteed to converge to a local maximum of the likelihood function.

Of course, for all the convergence results there must be persistence of excitation in the
models, (and estimators).

The second issue addressed in this chapter is whether schemes exist with the computational
complexity, of order $(N^2 T)$, of the Baum-Welch scheme, but with the on-line capability of
the recursive conditional-mean estimate scheme of Chapters 2 and 3 of [18], which is of order
$N^4 T$. To this end, known recursive schemes for transition probability, (or number), estimation
are generalized to achieve smoothed estimates, along with their associated approximations, where errors are mitigated by the smoothing.

In Section 4.2, the discrete-time HMM signal model to which the new theory applies is defined, measure changes are introduced, and associated convergence properties are studied. Parameter estimators are then proposed. In Section 4.3, the relevant conditional mean filters and smoothers are presented as on-line algorithms. In Section 4.4, model parameter and state estimation are studied for conditional mean estimates. Global convergence results are established showing asymptotic optimality and consistent estimation. Simulations are given in Section 4.5, and conclusions in Section 4.6.

4.2 Dynamics, Measure Change, and Martingale Properties

4.2.1 Dynamics

Our time parameter set is the non-negative integers $Z^+ = \{0, 1, 2, \ldots\}$. On a probability space $(\Omega, \mathcal{F}, P)$ we suppose we have a finite state, homogeneous Markov chain $\{X_\ell, \ell \in Z^+\$. As pointed out in [18], without loss of generality we can take the state space of the Markov chain to be the set $S = \{e_1, e_2, \ldots, e_n\}$ of unit vectors in $R^N$. Here $e_i = (0, 0, \ldots, 1, \ldots, 0)' \in R^N$.

Consequently, at each $\ell, X_\ell \in S$. Write

$$a_{ji} = P(X_\ell = e_j | X_{\ell-1} = e_i)$$

$$A = (a_{ji}), \quad 1 \leq i, j \leq N.$$

Now if the vector process $V$ is defined by putting $V_k := X_k - AX_{k-1}$ it follows that $V$ is a martingale increment:

$$E[V_{k+1} | X_k] = 0 \in R^N.$$

The Markov chain $X$ is not observed directly; rather there is a, (scalar), sequence of observations $y_1, y_2, \ldots$ such that

$$y_k = (C'X_{k-1}) + (D'X_{k-1})w_k. \quad (4.2.1)$$

Here, $w_k$ is a sequence of independent $N(0, 1)$ random variables defined on $(\Omega, \mathcal{F}, P)$. That $y_k$ is linear in $X_{k-1}$ is not really a restriction since nonlinear operations $f(X)$ are linear in
$X$ as with $f = (f(e_1), f(e_2), \ldots, f(e_N)), f(X) = (f, X)$. Therefore, $C = (c_1, \ldots, c_N)'$, $D = (d_1, \ldots, d_N)'$. Write

$$
\mathcal{G}_k = \text{the sigma field generated by } \{X_0, X_1, \ldots, X_k, y_1, \ldots, y_k\} \quad \text{and}
$$

$$
\mathcal{Y}_k = \text{the sigma field generated by } \{y_1, \ldots, y_k\}
$$

for the sigma-fields generated by the histories of the $X$ and $y$, and $y$ processes, respectively.

We shall write $M$ for the model determined by these parameters $(a_{ij}, c_i, d_i), 1 \leq i, j, \leq N$. Finally, we assume throughout that the model order, $N$, is known. The problem of estimating $N$ is beyond the scope of this chapter, see [25] for recent work in this area.

### 4.2.2 Measure Change

From [18] recall that a probability $\bar{P}$ is introduced such that, under $\bar{P}$, $X$ is still a Markov chain with transition matrix $A$, but the random variables $y_k$ are themselves independent and $N(0,1)$.

$$
\lambda_t(X_{t-1}) = \frac{\phi((y_t - C'X_{t-1})/(D'X_{t-1}))}{(D'X_{t-1})\phi(y_t)}
$$

where $\phi(x)$ is, for example $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

With $A_0 = 1$ and $\Lambda_k = \prod_{t=1}^{k} \lambda_t(X_{t-1})$, a probability measure $P$ can be defined on $(\Omega, \mathcal{G}_\infty)$ by putting $\frac{dP}{d\bar{P}}|_{\mathcal{G}_k} = \Lambda_k$.

One can then show (see [18]), that under $P$, $\{w_k\}$ is a sequence of independent $N(0,1)$ random variables, where $w_k := (y_k - (C'X_{k-1}))/((D'X_{k-1})$.

Furthermore, under $P$, $X$ remains a Markov chain with transition matrix $A$. That is, under $P$,

$$
X_k = AX_{k-1} + V_k  \quad (4.2.2)
$$

$$
y_k = (C'X_{k-1}) + (D'X_{k-1})w_k. \quad (4.2.3)
$$

However, $\bar{P}$ is a nicer measure with which to work. Write, $E$ (resp. $\bar{E}$), for expectation with respect to $P$, (resp. $\bar{P}$).

Let us denote the model (4.2.2), (4.2.3) as

$$
M = M(A, C, D, \pi_0)
$$
where \( \pi_0 = E[X_0] \).

We close this section by recalling a conditional version of Bayes’ theorem (see for example, Theorem 3.2 of Chapter 2 in [18]). Suppose \( \phi \) is a \( \mathcal{G}_k \)-measurable integrable random variable. Then

\[
E[\phi \mid y_k] = \frac{E[\mathbb{A}_k \phi \mid y_k]}{E[\mathbb{A}_k \mid y_k]} 
\]  

(4.2.4)

### 4.2.3 Transitions and Occupation Times

For \( 1 \leq r, s \leq N \), write \( J_{k}^{rs} = \sum_{t=1}^{k} \langle X_{t-1}, e_r \rangle \langle X_t, e_s \rangle \), then \( J_{k}^{rs} \) is the number of jumps of \( X \) from state \( r \) to state \( s \) up to time \( k \). The Markov chain \( X \) is not observed directly, but only through the observations \( j \). In Chapter 3 of [18] a recursive filter is given for an estimate of the vector process \( J_{k}^{rs}X_k \) for each \( r, s, 1 \leq r, s \leq N \). Updates of this vector quantity requires \( N^2 \) multiplications at each time \( k \). The occupation times for being in state \( r \) up until time \( k \) are \( \mathcal{O}_k^r = \sum_{s=1}^{k} J_{k}^{rs} \).

Let us denote

\[
J_k := (J_k^{rs}), \quad 1 \leq r, s \leq N \in R^{N \times N} 
\]

(4.2.5)

\[
\mathcal{O}_k := J_k^{1'} \in R^{N} 
\]

(4.2.6)

where \( 1' \) denotes the row vector \([1, 1, \ldots, 1]\). Thus in obvious notation

\[
J_k = \sum_{t=1}^{k} X_{t-1}X_t', \quad (\mathcal{O}_k)_{\text{diag}} = \sum_{t=1}^{k} X_{t-1}X_t' = \sum_{t=1}^{k} (X_{t-1})_{\text{diag}}. 
\]

(4.2.7)

(4.2.8)

Post multiplication of (4.2.2) by \( X_{k-1}' \) and summing yields

\[
k^{-1}J_k' = A(k^{-1} \mathcal{O}_k)_{\text{diag}} + k^{-1} \sum_{t=1}^{k} V_t X_{t-1}'. 
\]

(4.2.9)

Given knowledge of \( \{X_k\} \), it makes sense to estimate \( A \) as

\[
\tilde{A}_k = J_k'(\mathcal{O}_k)_{\text{diag}}^{-1} 
\]

(4.2.10)

at least when the inverse exists (as when all states in \( \{X_k\} \) are persistently excited). We say the states in the state sequence \( \{X_k\} \), which is a particular outcome of the HMM, are persistently excited when there is a finite integer \( M > 0 \), such that \( O_k^r > 0 \) if \( k > M \) for
\[ i = 1, \ldots, N, \text{and a constant } B < \infty \text{ such that for all } k > M \text{ and } 1 \leq i \leq N \text{ then } \frac{k}{k} < B. \]

This is equivalent to the condition that

\[
\lim_{k \to \infty} \sup_{k} \left\{ \frac{1}{k} \mathcal{O}_{k} \right\}_{\text{diag}}^{-1} < \infty.
\] (4.2.11)

In a parallel manner to the above, one can define transitions \( T_{k}^{\prime} (f) = \sum_{\ell=1}^{k} (X_{\ell-1}, e_{\ell}) f(y_{\ell}) \)
where \( f(y) \) is either \( y \), or \( y^{2} \), depending on application. So define the row vector with elements \( T_{k}^{\prime} \) as

\[
T_{k}^{\prime} (f) = \sum_{\ell=1}^{k} f(y_{\ell}) X_{\ell-1}^{\prime}.
\] (4.2.12)

Now (4.2.3) leads to

\[
k^{-1} T_{k}^{\prime} (y) = C' k^{-1} (O_{k})_{\text{diag}} + D' k^{-1} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} w_{\ell} \]

and thus estimates can be defined as

\[
\tilde{C}_{k} = T_{k}^{\prime} (O_{k})_{\text{diag}}^{-1}.
\] (4.2.13)

Likewise, squaring (4.2.3) and post-multiplying by \( X_{k-1}^{\prime} \) we have:

\[
\frac{1}{k} T_{k}^{\prime} (y^{2}) = (c_{1}^{2}, c_{2}^{2}, \ldots) \frac{1}{k} (O_{k})_{\text{diag}} + (d_{1}^{2}, d_{2}^{2}, \ldots) \frac{1}{k} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} w_{\ell}^{2}
\]

\[
+ 2(c_{1} d_{1}, c_{2} d_{2}, \ldots) \frac{1}{k} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} w_{\ell}.
\] (4.2.15)

Estimates \( \tilde{D} \) can be constructed from

\[
(d_{1}^{2}, d_{2}^{2}, \ldots, d_{N}^{2}) = T_{k}^{\prime} (y^{2}) (O_{k})_{\text{diag}}^{-1} - (c_{1}^{2}, c_{2}^{2}, \ldots, c_{N}^{2}).
\] (4.2.16)

The above estimators (4.2.10), (4.2.14) and (4.2.16) have a similar form to the Baum-Welch EM algorithm for estimating hidden Markov model parameters, see [68]. However, unlike the off-line EM algorithm, these estimators can be used on-line as new data points are received.

We now proceed to develop almost sure convergence properties using a martingale convergence theorem and the Kronecker lemma [53, 64]. Also, mean square convergence is studied using martingale properties to achieve mildly stronger convergence rates.
4.2.4 Strong Convergence

We first introduce a function $\rho(k)$ satisfying the requirement that $\sum_{k=1}^{\infty} \rho(k)^{-1}$ exists. For example, define $\rho(k) = k^2 \bar{\rho}(k)$ where $\bar{\rho}(k)$ is defined as

$$
\bar{\rho}(k) = \begin{cases} 
\frac{\ln^4(\ln k)^\alpha}{k} & k \in [1, 4) \\
\frac{\ln^4(\ln k)^\alpha}{k} & k \in [4, \infty)
\end{cases}
$$

for $\alpha > 1$. The following lemma now holds

**Lemma 4.2.1** Consider the HMM defined by (4.2.2)/(4.2.3) and a particular observation sequence and state sequence outcome, $\{y_k\}$ and $\{X_k\}$, of the HMM where states $X_k$ are $\mathcal{Y}_k$-measurable and all states excited at least at some finite time instant, so that $\lim_{k \to \infty} (O_k)_{\text{diag}}^{-1}$ exists. Then $\lim_{k \to \infty} \sup \bar{A}_k, \bar{C}_k, \bar{D}_k$ exists almost surely. Moreover, under the weak excitation conditions $\lim_{k \to \infty} (O_k)_{\text{diag}}^{-1} = 0$, and the extra structural condition $1 > a_{ij} > 0 \forall i, j$ then

$$
\lim_{k \to \infty} \bar{A}_k, \bar{C}_k, \bar{D}_k = A, C, D \ a.s.
$$

Furthermore, in the absence of the above structural condition, but under the persistence of excitation condition (4.2.11), the almost sure convergence rate is guaranteed to be $\bar{\rho}(k)^{-\frac{1}{2}}$ which is like $\frac{1}{\sqrt{k}} (\ln k)^{\frac{\alpha}{2}}$ for $k > 4$ and any $\alpha > 1$, and the mean square convergence rate is guaranteed to be arbitrarily slower than $k^{-\frac{1}{2}},$ denoted $(k^{-\frac{1}{2}})$.

**Proof:** We first proceed with the almost sure convergence rate results, under the persistence of excitation condition.

Now $S_k := \sum_{t=1}^{k} \rho(k)^{-\frac{1}{2}} V_t X_t^i X_t^j$ is a matrix whose elements are martingales on $(P, \mathcal{G}, \Omega)$, since $E \left[ S_k^{ij} | \mathcal{G}_{k-1} \right] = S_k^{ij}$ for each $i, j$, where $S_k^{ij}$ denotes the $ij$-th component of $S_k$. Also $S_k^{ij}$ is bounded in $L^2$ for each $i, j$ since

$$
E[(S_k^{ij})^2] = E \left[ \sum_{s=1}^{k} \sum_{r=1}^{k} \rho(s)^{-\frac{1}{2}} V_s^i V_s^j X_s^i X_s^j \rho(r)^{-\frac{1}{2}} \right]
$$

$$
= E \left[ \sum_{s=1}^{k} \sum_{r=1}^{k} E \left[ \rho(s)^{-\frac{1}{2}} V_s^i V_s^j X_s^i X_s^j \rho(r)^{-\frac{1}{2}} \big| \mathcal{G}_{\max(t, r - 1)} \right] \right]
$$

$$
= E \left[ \sum_{r=1}^{k} \rho(r)^{-1} (X_r^i X_r^j)^2 E \left[ (V_r^i)^2 \big| \mathcal{G}_{t-1} \right] \right]
$$

$$
< K \sum_{r=1}^{k} \rho(r)^{-1} \text{ for some } K
$$

$$
< \infty \text{ for each } i, j. \tag{4.2.17}
$$
Here we have used the readily established property,

\[ E [V_k'V_k \mid \mathcal{G}_{k-1}] = 1 - tr(A'AX_{k-1})_{\text{diag}} \leq K \]

for some \( K < \infty \), and the properties that \( E [X_k \mid \mathcal{G}_{k-1}] = AX_{k-1} \) and that \( E[V_k \mid \mathcal{G}_{k-1}] = 0 \).

Thus, by martingale convergence, [64, 59]

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{\ell=1}^{k} \rho(\ell)^{-\frac{1}{2}} V_{\ell}X_{\ell-1}' = 0 \quad \text{a.s.} \tag{4.2.18}
\]

Using the Kronecker lemma [53, 64], since \( \rho(k)^{-\frac{1}{2}} \to 0 \) as \( k \to \infty \),

\[
\lim_{k \to \infty} \phi_k = 0 \quad \text{a.s.}
\]

where \( \phi_k = \rho(k)^{-\frac{1}{2}} \sum_{\ell=1}^{k} V_{\ell}X_{\ell-1}' \). \tag{4.2.19}

Here 0 denotes the matrix with all zero elements. Indeed \( \frac{1}{k} \sum_{\ell=1}^{k} V_{\ell}X_{\ell-1}' \to 0 \) a.s. at a rate \( \bar{p}(k)^{\frac{1}{2}} \), which is like \( \frac{1}{\sqrt{k}}(ln(k \cdot ln(k))^\alpha)^{\frac{1}{2}} \) for \( k > 4 \) and any \( \alpha > 1 \).

Under the persistently exciting assumption (4.2.11), and using the result that a limit of products is the product of limits (when these exist), then

\[
\bar{A}_k := \bar{A}_k - A = \sum_{\ell=1}^{k} V_{\ell}X_{\ell-1}'[O_k]_{\text{diag}}^{-1}
\]

\[
= \left( \frac{1}{k} \sum_{\ell=1}^{k} V_{\ell}X_{\ell-1}' \right) \left( \frac{1}{k} O_k \right)_{\text{diag}}^{-1}
\]

\[
\to 0 \quad \text{a.s.}
\]

at a rate \( \bar{p}(k)^{\frac{1}{2}} \). A result holds similarly for \( \bar{C}_k := \bar{C}_k - C \).

To achieve corresponding results for \( \bar{D}_k \), we replace \( V_{\ell} \) by \( w_{\ell} \). Recalling \( E [w_{\ell} \mid \mathcal{G}_{\ell-1}] = 0 \), it is not surprising then that

\[
\frac{1}{k} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} w_{\ell} \to 0 \quad \text{a.s.} \quad k \to \infty
\]

at the rate \( \bar{p}(k)^{\frac{1}{2}} \).

From (4.2.15) two martingales must be considered:

\[
2(c_1d_1, \ldots, c_Nd_N)\frac{1}{k} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} w_{\ell} \tag{4.2.20}
\]

\[
(d_1^2, \ldots, d_N^2)\frac{1}{k} \sum_{\ell=1}^{k} (X_{\ell-1})_{\text{diag}} [w_{\ell}^2 - 1]. \tag{4.2.21}
\]
These martingales again converge almost surely, so Kronecker’s lemma can be applied as before.

Now we consider the the first result of the lemma, where \( \lim_{k \to \infty} (O_k)^{-1}_{\text{diag}} \) exists.

Let us write the estimation error as, \( \tilde{A}_k = (\sum_{l=1}^k V^*_l X^*_l) (O_k)^{-1}_{\text{diag}} \). Because \( O_k \) is diagonal then the \( ij \)th element of \( \tilde{A}_k \) can be written as,

\[
\tilde{A}^{ij}_k = (\sum_{l=1}^k V^*_l X^*_l) \frac{1}{O^*_k}.
\]  

(4.2.22)

To prove the first lemma result, consider that each state is perhaps not excited (a state \( i \) is excited when \( \lim_{k \to \infty} (O^*_k)^{-1} = 0 \)).

Consider the two cases of excited and non-excited states separately. For the states denoted \( m \) which are excited, then

\[
|\tilde{A}^{im}_k| = \left| \sum_{l=1}^k V^*_l X^*_l \right| \leq \max_l |V^*_l| \sum_{l=1}^k X^*_l = \max_l |V^*_l|.
\]

Hence \( \lim_{k \to \infty} \sup_i |\tilde{A}^{im}_k| \) exists \( \forall i \) for states \( m \) that are excited.

For all the states denoted \( j \) which are not excited, then \( \lim_{k \to \infty} \sum_{l=1}^k V^*_l X^*_l \) exists \( \forall i \), because \( X^*_l = 0, \forall \ell > T \) for some \( T \) and likewise \( \lim_{k \to \infty} O^*_k \) exists. Hence, \( \lim_{k \to \infty} \tilde{A}^{ij}_k \) exists \( \forall i \) for states \( j \) that are not excited. Hence, under the excitation condition that \( \lim_{k \to \infty} (O^*_k)^{-1} \) exists then

\[
\lim_{k \to \infty} \sup_k \tilde{A}_k \text{ exist a.s.}
\]

For convergence under the weaker excitation condition \( \lim_{k \to \infty} (O_k)^{-1}_{\text{diag}} = 0 \), note that \( \varphi^{ij}_k = \sum_{l=1}^k V^*_l X^*_l \) is a martingale on \( G_{k-1} \), since \( E[\varphi^{ij}_k | G_{k-1}] = \sum_{l=1}^{k-1} V^*_l X^*_l = \varphi^{ij}_{k-1} \). The scalar process variance \( \Phi^{ij}_k \), of \( \varphi^{ij}_k \), is given by

\[
\Phi^{ij}_k := \sum_{l=1}^k E[(\varphi^{ij}_l - \varphi^{ij}_{l-1})^2 | G_{l-1}]
\]

\[
= \sum_{l=1}^k E[(V^*_l X^*_l)^2 | G_{l-1}]
\]

\[
= \sum_{l=1}^k X^*_l E[(V^*_l)^2 | G_{l-1}].
\]

Now \( E[(V^*_l)^2 | G_{l-1}] = \sum_{j=1}^N (a_{ij} - a^2_{ij})X^*_l \), hence \( K_U \geq E[(V^*_l)^2 | G_{l-1}] \geq K_L \), for some \( K_U, K_L > 0 \). The extra structural condition \( 1 > a_{ij} > 0 \) is required to find a positive lower
bound. It follows that

\[
\left( \sum_{t=1}^{k} X^j_{t-1} \right) K_U \geq \Phi^j_k \geq \left( \sum_{t=1}^{k} X^j_{t-1} \right) K_L
\]

(4.2.23)

for some \( K_U, K_L > 0 \). Now under the weak excitation condition, \( \lim_{k \to \infty} (C_k)^{-1}_{\text{diag}} = 0 \), the process variance is unbounded, i.e. \( \lim_{k \to \infty} \Phi^j_k = \infty \). In [59] it is shown that a scalar martingale normalized by its process variance, which is unbounded in the limit, converges to zero, that is,

\[
\lim_{k \to \infty} \frac{\varphi_k^j}{\Phi_k^j} = 0.
\]

(4.2.24)

The first part of the lemma follows by noting that \( \Phi^j_k \) is lower bounded by \( (C_k^j)_{k-1}^1 \). Then rewriting \( \tilde{A}^j_k \) as follows

\[
|\tilde{A}^j_k| = \left| \frac{\varphi_k^j}{(C_k^j)_{k-1}} \right| \leq \left| \frac{\varphi_k^j}{\Phi_k^j} \right| \frac{1}{K_L}
\]

and combined with the (4.2.24) convergence result leads to the result stated in the lemma.

That is,

\[
\lim_{k \to \infty} \tilde{A}_k = A \quad \text{a.s.}
\]

The first part of the lemma follows as claimed, at least for \( \tilde{A}_k, \tilde{C}_k \) (with similar extra structural condition to give convergence under the weak excitation condition). The corresponding result for \( \tilde{D}_k \) holds by a similar line of reasoning by replacing \( V_k \) with \( w_k \) above.

For mean square convergence, by using martingale properties (as above), \( \tilde{S}_k := \sum_{t=1}^{k} V_t X^j_{t-1} \) is a martingale on \( (P, G, \Omega) \) and with a covariance growing (in all its eigenvalues) at a rate \( k \).

Equivalently letting \( k \), denote a rate arbitrarily faster than \( k \), we have

\[
\lim_{k \to \infty} (k_+)^{-1} E \left[ \tilde{S}_k \tilde{S}_k^t \right] = 0,
\]

and thus \( (k_+)^{-\frac{1}{2}} \sum_{t=1}^{k} V_t X^j_{t-1} \to 0 \) in mean square. Consequently \( \tilde{A}_k \to 0 \) in mean square at a rate \( (k_-)^{-\frac{1}{2}} \), as do \( \tilde{C}_k, \tilde{D}_k \).

Remark:

1. The extra structural condition, \( 1 > a_{ij} > 0 \), required to prove convergence under the weak excitation condition is a sufficient but possibly not necessary condition.

2. The maximum of the likelihood function is achieved at the true values of the HMM parameters, see [46], hence the estimators given here locate the global optimum of the likelihood surface almost surely.
Before proceeding to consider parameter estimation when the state sequence \( \{X_k\} \) is not measured directly we need to develop conditional mean filters for various quantities.

### 4.3 Conditional Mean Filters and Smoothers

Let us denote associated conditional mean estimates based on the correct model and a sequence of estimated models as

\[
\hat{X}_{k|k} = E\left[X_k \mid \mathcal{Y}_k, \mathcal{M}\right]; \quad \hat{X}_{k|k}, \hat{X}_{k|\ell} = E\left[X_k \mid \mathcal{Y}_k, \hat{\mathcal{M}}_k\right]
\]

\[
\hat{J}_{k|k} = E\left[J_k \mid \mathcal{Y}_k, \mathcal{M}\right]; \quad \hat{J}_{k|k}, \hat{J}_{k|\ell} = E\left[J_k \mid \mathcal{Y}_k, \hat{\mathcal{M}}_k\right]
\]

\[
\hat{O}_{k|k} = E\left[O_k \mid \mathcal{Y}_k, \mathcal{M}\right]; \quad \hat{O}_{k|k}, \hat{O}_{k|\ell} = E\left[O_k \mid \mathcal{Y}_k, \hat{\mathcal{M}}_k\right]
\]

\[
\hat{\tau}_{k|k}(f) = E\left[\tau_k(f) \mid \mathcal{Y}_k, \mathcal{M}\right]; \quad \hat{\tau}_{k|k}, \hat{\tau}_{k|\ell}(f) = E\left[\tau_k(f) \mid \mathcal{Y}_k, \hat{\mathcal{M}}_k\right]
\]

(4.3.1)

where \( \hat{\mathcal{M}}_k \) denotes a sequence of model estimates, that is, \( \hat{\mathcal{M}}_k = \hat{\mathcal{M}}_1, \hat{\mathcal{M}}_2, \ldots, \hat{\mathcal{M}}_k \), where \( \hat{\mathcal{M}}_k \) is the estimated model at time instant \( k \). Similar notation will denote one step ahead predictions \( \hat{X}_{k,k-1} \) etc., or smoothed estimates \( \hat{X}_{k,k+d} \) etc.

#### 4.3.1 On-Line Algorithms

When seeking on-line estimates of \( \mathcal{J}_k \), such as conditional mean filtered or smoothed estimates, it turns out, perhaps surprisingly, to be essential to work first with on-line estimates of \( \mathcal{J}_k \) given \( X_k = e_i \) for each \( i \) and then derive the desired estimates from these. Thus define

\[
\mathcal{J}_k^X := X_k \text{ row vec } \mathcal{J}_k \in \mathbb{R}^{N \times N^2}
\]

(4.3.2)

\[
\text{ row vec } \mathcal{J}_k = 1' \mathcal{J}_k^X \in \mathbb{R}^{1 \times N^2}.
\]

(4.3.3)

Let us work again with conditional mean estimates for filtering and smoothing, denoted as follows for \( 0 \leq k \leq T \). In the first instance, we condition on knowledge of a known model \( M \) as in [18], but later point out that we could also work with some model estimate \( \hat{\mathcal{M}}_k \). Thus define

\[
\hat{\mathcal{J}}_{k|k} := E\left[\mathcal{J}_k \mid \mathcal{Y}_k, \mathcal{M}\right], \quad \hat{\mathcal{J}}_{k|T} := E\left[\mathcal{J}_k \mid \mathcal{Y}_T, \mathcal{M}\right].
\]

(4.3.4)

\( \hat{J}_{k|k}^X, \hat{J}_{k|T}^X, \hat{X}_{k|k}, \hat{X}_{k|T} \) are defined similarly. Also, denote associated unnormalized forms under \( \mathbb{P} \)

\[
\sigma(\mathcal{J}_{k|k}) := \mathbb{E}[\Lambda_k \mathcal{J}_k \mid \mathcal{Y}_k, \mathcal{M}], \quad \sigma(\mathcal{J}_{k|T}) := \mathbb{E}[\Lambda_T \mathcal{J}_k \mid \mathcal{Y}_T, \mathcal{M}]
\]

(4.3.5)
and likewise for \( \sigma(\mathcal{J}_k^X) \), \( \sigma(X_k|\mathcal{Y}_T) \). Actually, it is more conventional to denote \( \sigma(X_k|T) = \alpha_k \), \( \sigma(X_k|T) = \gamma_k|T \) and define a vector \( \beta_k|T \) with ith element

\[
\beta_k(i)|T = \tilde{E} \left[ \Lambda_{k+1,T} | \mathcal{Y}_T, X_k = e_i \right], \quad \Lambda_{k+1,T} = \prod_{t=k+1}^T \lambda_t(X_{t-1}). \tag{4.3.6}
\]

Clearly, here (as is known)

\[
\gamma_k|T = \text{diag} \alpha_k \beta_k|T = \text{diag} \beta_k|T \alpha_k \tag{4.3.7}
\]

This gives:

\[
\tilde{E} \left[ \Lambda_{1,T} | \mathcal{Y}_T, M \right] = \langle \alpha_T, 1 \rangle = \langle \beta_0|T, 1 \rangle \tag{4.3.8}
\]

or, as is shown perhaps more easily after Theorem 4.4.1, \( \tilde{E} \left[ \gamma_0,T | \mathcal{Y}_T \right] = \langle \alpha_k, \beta_k|T \rangle \) independent of \( k \).

We derive smoothed estimates of \( \mathcal{J}_k, \mathcal{J}_k^X \) in terms of filtered estimates and \( \beta \) in the following lemma.

**Lemma 4.3.1**

\[
\sigma(\mathcal{J}_k^X|T) = (\beta_k|T)_{\text{diag}} \sigma(\mathcal{J}_k^X) \tag{4.3.9}
\]

row vec \( \sigma(\mathcal{J}_k|T) = \langle \beta_k|T, \sigma(\mathcal{J}_k^X) \rangle \tag{4.3.10} \]

**Proof:** Denoting \( \mathcal{Y}_k|T = \sigma(y_k, y_{k+1}, \ldots, \mathcal{Y}_T) \), then

\[
\sigma(\mathcal{J}_k^X|T) = \tilde{E} [\Lambda_T X_k \ \text{row vec} \ \mathcal{J}_k | \mathcal{Y}_T, M] = \tilde{E} [\Lambda_k X_k \ \text{row vec} \ \mathcal{J}_k \tilde{E} [\Lambda_{k+1,T} | \mathcal{Y}_T, \mathcal{J}_k, M] | \mathcal{Y}_T, M]
\]

where \( \mathcal{J}_k = \sigma(X_1, X_2, \ldots, X_k) \). Now since \( X_k \) is Markov, \( \mathcal{J}_k \) can be replaced by \( X_k \). Also, since nonlinear operations on \( X \in \{e_1, e_2, \ldots, e_N\} \) are linear in \( X \) as

\[
f(X) = \langle f(e_1), f(e_2), \ldots, f(e_N) \rangle, X \rangle,
\]

then

\[
\sigma(\mathcal{J}_k^X) = \sum_{i=1}^N \tilde{E} [\Lambda_k X_k^i \ \text{row vec} \ \mathcal{J}_k \beta_k|T, e_i | \mathcal{Y}_T, M]
\]

as the \( y_i \) are i.i.d., \( \mathcal{Y}_T \) can be replaced by \( \mathcal{Y}_k \) to yield the desired result.

To recall the well known recursions for \( \alpha_k, \beta_k|T \), we define

\[
B(y_t) = \Gamma(y_t)_{\text{diag}} \tag{4.3.11}
\]
where the \( N \)-vector \( \Gamma(y_t) \) has \( i \)th element
\[
\Gamma^i(y_t) = \lambda(e_i) = \frac{\phi((y_t - c_i)/d_i)}{d_i \phi(y_t)}.
\]

It is readily verified that
\[
\alpha_k = AB(y_k)\alpha_{k-1}, \quad \alpha_0 = \pi_0 \quad (4.3.12)
\]
\[
\beta_k\|T = B(y_k)A'\beta_{k+1}\|T, \quad \beta_T = 1
\]

The following lemma is a convenient re-packaging of the results of [18, Chapter 3] to a matrix form.

**Lemma 4.3.2** Consider the HMM signal model (4.2.2) (4.2.3) denoted by \( M \). The conditional mean filtered estimates for \( J_k^X, J_k, T_k^X, T_k \), under \( \tilde{P} \), are obtained from:
\[
\sigma(J_k^X) = AB(y_k)\sigma(J_{k-1}^X|k-1) + \{ (Ae_1)_{\text{diag}}, (Ae_2)_{\text{diag}}, \ldots, (Ae_N)_{\text{diag}} \} (B(y_k)(\alpha_{k-1})_{\text{diag}} \otimes I)
\]
row vec \( \sigma(J_k^X) = \begin{bmatrix} \sigma(J_{k}^X) \end{bmatrix} \) (4.3.14)
\[
\sigma(T_k^X) = AB(y_k)\sigma(T_{k-1}^X|k-1) + AB(y_k)(\alpha_{k-1})_{\text{diag}}f(y)
\]
row vec \( \sigma(T_k^X) = \begin{bmatrix} \sigma(T_k^X) \end{bmatrix} \).

**Proof:** Now
\[
\sigma(J_k^X) = \tilde{E} [\Lambda_{k-1}\lambda_k(X_{k-1})X_k \text{ row vec } (J_{k-1} + X_{k-1}X_k') | \mathcal{Y}_k, M].
\]

In the following manipulations, when appropriate, we replace \( X_k \) by \( (AX_k + V_k) \) and exploit the property \( \tilde{E} [\Lambda_k V_k | \mathcal{Y}_k, M] = 0 \), since \( X \) is independent of \( y \) under \( \tilde{P} \). Also, we recall \( X_k, X_k' = (X_k)_{\text{diag}} \) and, more generally, as already noted, \( f(X) = \langle f(e_1), f(e_2), \ldots, f(e_N) \rangle, X \).

Then
\[
\sigma(J_k^X) = \tilde{E} [\Lambda_{k-1}AB(y_k)X_{k-1} \text{ row vec } J_{k-1}|k-1 | \mathcal{Y}_k, M]
\]
\[
+\tilde{E} [\Lambda_{k-1} \lambda_k(X_{k-1})X_{k-1}' \otimes (X_k)_{\text{diag}} | \mathcal{Y}_k, M]
\]
\[
= \tilde{E} [\Lambda_{k-1}AB(y_k)J_{k-1}^X | \mathcal{Y}_k, M]
\]
\[
+\tilde{E} [\Lambda_{k-1} (B(y_k)X_{k-1})' \otimes (AX_{k-1})_{\text{diag}} | \mathcal{Y}_k, M].
\]

Because the \( y_k \) are i.i.d. \( \mathcal{Y}_k \) can be replaced by \( \mathcal{Y}_{k-1} \) leading to the result. (Here \( A \otimes B \) denotes the Kronecker product, with \( (A \otimes B)_{ij} = a_{ij}B \).)
A similar discussion for \( \sigma(T_{k|k}^X) \) gives

\[
\sigma(T_{k|k}^X) = \bar{E} \left[ \Lambda_{k-1} \lambda_k (X_{k-1}) X_k \right] \text{ row vec } \left( T_{k|k} + X_{k-1} f(y_k) \right) | \mathcal{Y}_k, M \]
\[
= \bar{E} \left[ \Lambda_{k-1} AB(y_k) T_{k-1}^X | \mathcal{Y}_k, M \right]
+ \bar{E} \left[ \Lambda_{k-1} B(y_k)(X_{k-1})_{\text{diag}} f(y_k) | \mathcal{Y}_k, M \right].
\]

The desired result follows. \( \square \)

**Remarks:**

1. Exponential stability or initial-condition forgetting of the filters (4.3.14) and (4.3.15) follows by appealing to the generalized Perron-Frobenius result [69] in the same way as in [70]

2. The computational difficulty is that the \( \sigma(T_{k|k}^X) \) calculation requires \( N^4 \) multiplications for each update.

3. If the observations have zero-delay, so that \( y_k = (C'X_k) + (D'X_k w_k) \) then
\[
\lambda_k = \frac{\Phi((y_k - C'X_k)/(D'X_k w_k))}{\Phi(y_k)}
\]
and the updating formula effectively interchanges the \( B \) and \( A \) matrices so that:

\[
\alpha_k = B(y_k) A \alpha_{k-1}
\]

\[
\sigma(J_{k|k}^X) = B(y_k) A \sigma(J_{k|k}^X) + \left[ (B(y_k) A e_1)_{\text{diag}} \cdots \right] ((\alpha_{k-1})_{\text{diag}} \otimes I)
\]

4. Note \( O_k = J_{k|k} \), so that \( \sigma(O_{k|k}) = \sigma(J_{k|k}) \). (Note also that \( \alpha_k \) can be derived from \( \sigma(J_{k|k}^X) \) by summation operations).

5. The derivations and results stated above have been developed for time-homogeneous models \( M \), so conditioning is on \( M \). However, they also apply mutatis mutandis with appropriate substitutions for estimates of \( M \), including time-varying estimates such as can be derived on-line. Thus, for the case when \( A \) is being estimated as here, \( A \) in the recursions for \( \sigma^*(\cdot) \) at time \( k \) can be replaced by \( \bar{A}_{k-1} \) (for example). Then \( M \) is replaced in the derivations by \( \bar{M}_k = [\bar{M}, \bar{M}, \ldots, \bar{M}] \) where \( \bar{M}_k = (\bar{A}_{k-1}, C, D, \pi_0) \).

Of course, the estimates should then be denoted \( \sigma(J_{k|k}^X, \bar{M}_k) \) etc. Again asymptotically optimum state estimation is achieved under the relevant excitation conditions.
6. Under the partially justified assumption that \( X_k \) is uncorrelated with \( \mathcal{J}_k \), a suboptimal version of the above results is readily obtained. More precisely, “assume” that

\[
\sigma(\mathcal{J}_k^X) = \hat{X}_{k|k} \quad \text{row vec} \quad \sigma(\mathcal{J}_k^X)
\]

then simple manipulations yield

\[
\sigma(\mathcal{J}_k^X) = \sigma(\mathcal{J}_{k-1|k-1}) + (\alpha_{k-1})_{\text{diag}} B(y_k) A'
\]

requiring computational effort \( O(N)^2 \). Actually, this is equivalent to omitting the backward pass in the Baum-Welch procedure. Versions involving smoothed estimates are now developed.

### 4.3.2 Algorithms Based on Smoothing

Let us define the transitions

\[
\xi_k := X_{k-1} X'_{k}, \quad \xi_k^X := X_k \quad \text{row vec} \xi_k
\]

Then, as above we can show that

\[
\tilde{\xi}_{k|k}^X := E \left[ \Lambda_k \xi_k^X \mid \mathcal{Y}_k \right] = \left( (A e_1)_{\text{diag}} \cdots (A e_N)_{\text{diag}} \right) (B(y_k) \alpha_{k-1} \otimes I) .
\]

Indeed

\[
\sigma(\mathcal{J}_{k|k}^X) = E \left[ \Lambda_k \mathcal{J}_{k|k}^X \mid \mathcal{Y}_k \right] = \sum_{\ell=1}^k E \left[ \Lambda_k \xi_k^X \mid \mathcal{Y}_k \right]
\]

\[
= AB(y_k) AB(y_{k-1}) \cdots AB(y_1) \xi_1^X
\]

\[
+ AB(y_k) \cdots AB(y_2) \xi_2^X
\]

\[
\cdots + \tilde{\xi}_{k|k}^X . \tag{4.3.17}
\]

Furthermore, row vec \( \mathcal{J}_k = 1' \mathcal{J}_k^X \), so that a reformulation of the Baum-Welch algorithm is

\[
\text{row vec} \quad \sigma(\mathcal{J}_{k|k}) = \left\langle B(y_1) A' B(y_2) A' \cdots B(y_k) A' \left( \left[ \xi_1^X \right]_1 \right) \right\rangle
\]

\[
+ \left\langle B(y_2) A' B(y_3) A' \cdots B(y_k) A' \left( \left[ \xi_2^X \right]_2 \right) \right\rangle
\]

\[
+ \cdots + \left\langle 1' \xi_{k|k} \right\rangle .
\]

We consider a modification with \( \sigma(\mathcal{J}_{k|k}^X) := E \left[ \Lambda_k \mathcal{J}_{k|k}^X \mid \mathcal{Y}_k \right] \) as

\[
\tilde{\mathcal{J}}_{k+p|k+p}^X = \tilde{E} \left[ \Lambda_{k+p} \mathcal{J}_{k+p}^X \mid \mathcal{Y}_{k+p} \right]
\]
\[
\begin{align*}
\text{vec} & = AB(y_{k+p})AB(y_{k+p-1}) \cdots AB(y_{k+1})\sigma(J_{k}^X) \\
& + AB(y_{k+p}) \cdots AB(y_{k+2})\tilde{\xi}_{k+1|k+1} \\
& \quad \cdots + \tilde{\xi}_{k+p|k+p}.
\end{align*}
\]

Consequently, since row vec \(\sigma(J_{k+p|k+p}) = 1^\prime \sigma(J_{k+p|k+p})\) and recalling (4.3.13),
\[
\text{row vec } \sigma(J_{k+p|k+p}) = \left( \beta_{k+1|k+p}, \sigma(J_{k|k}) \right) \\
+ \left( \beta_{k+2|k+p}, \tilde{\xi}_{k+1|k+1} \right) + \cdots + \left( 1, \tilde{\xi}_{k+p|k+p} \right). \quad (4.3.19)
\]

This in turn leads to a practical suboptimal scheme of \(O(N^2)\) by using suboptimal estimates for \(\sigma(J_{k|k})\) via (4.3.16). As \(p\) increases, the suboptimal estimates are known to become optimal (exponentially) via results in [8].

Other smoothing formulas can be derived for on-line estimation of the model parameter vector \(\pi_0\), being the a priori \(E[X_0]\). Equivalently, one can achieve on-line smoothed estimates of \(X_0\) as \(\hat{X}_{0|k,M} = E[X_0 | \mathcal{Y}_k, M]\), or more generally \(\hat{X}_{\ell|k,x_k}\) for \(k > \ell\). The approach is to work first with on-line estimates of
\[
X^X_{\ell|k} := X_k X^\ell_k, \quad X^\ell_k = 1^\prime X^X_{\ell|k}.
\]

We achieve recursions such as
\[
\sigma(X^X_{0|k}) := E \left[ \Lambda_k X^X_{0|k} \mid \mathcal{Y}_k, M \right] = AB(y_k)\sigma(X^X_{0|k-1}) \\
\sigma(X^\ell_{0|k}) := 1^\prime \sigma(X^X_{0|k}) \quad (4.3.21)
\]

which when normalized give the desired estimates \(\hat{X}^X_{0|k}, \hat{X}^\ell_{0|k}\).

Notice that these on-line formulas given the same results as the off-line Baum Welch scheme since from (4.3.21), \(\sigma(X^\ell_{0|k}) = \langle \beta_{0|k}, (\pi_0)_{\text{diag}} \rangle\).

We now proceed to consider parameter estimation based on these conditional mean estimates.

### 4.4 Estimation using Conditional Mean Estimates

In this section we consider the parameter estimates
\[
\hat{A}_{k|k,x_k} = \hat{J}^\prime_{k|k,x_k} \left( \hat{\sigma}_{k|k,x_k} \right)^{-1}_{\text{diag}} \quad \text{and}
\hat{C}_{k|k,x_k} = \hat{J}^\prime_{k|k,x_k} (f) \left( \hat{\sigma}_{k|k,x_k} \right)^{-1}_{\text{diag}}, \quad (4.4.1)
\]
and likewise for $\bar{D}_{k|k,\bar{X}_{t_k}}$ via (4.2.15).

Initially, we consider the (somewhat artificial) special case when it is assumed that the conditional mean estimates based on the true model are available, then $\bar{M}_{k+1} = M$ for all $k$.

Otherwise, the model estimate is adaptively updated as follows

$$\bar{M}_{k+1} = M(\bar{A}_{k|k,\bar{X}_{t_k}}, \bar{C}_{k|k,\bar{X}_{t_k}}, \bar{D}_{k|k,\bar{X}_{t_k}}, \bar{\pi}_0) \quad \text{and}$$

$$\bar{M}_{k+1} = \bar{M}_0, \ldots, \bar{M}_{k+1}. \quad (4.4.2)$$

We introduce the persistently excitation condition associated with the model $M$ and its "estimate" $\bar{M}$ as

$$\lim_{k \to \infty} \sup \left( \frac{1}{k} \bar{C}_{k|k,\bar{X}_{t_k}} \right)_{\text{diag}}^{-1} < \infty. \quad (4.4.3)$$

### 4.4.1 Towards Global Convergence

**Theorem 4.4.1** Consider the HMM of (4.2.2) (4.2.3) and a particular observation sequence and state sequence outcome, $\{y_k\}$ and $\{X_k\}$, of the HMM with all states in $\{X_k\}$ persistently exciting in that (4.2.11) holds. Consider the (somewhat artificial) case where conditional mean estimates based on the true model $M$ are known. Then

$$\lim_{k \to \infty} \bar{A}_{k|k,M}, \bar{C}_{k|k,M}, \bar{D}_{k|k,M} = A, C, D, \text{ a.s.} \quad (4.4.4)$$

The almost sure convergence rate guaranteed is $\bar{p}(k)^{\frac{1}{2}}$ which is like $\frac{1}{\sqrt{k}}(\ln k(\ln \ln k)^{\alpha})^{\frac{1}{2}}$, for $k > 4$ and for any $\alpha > 1$, and the mean square rate is $(k_-)^{-\frac{1}{4}}$.

**Proof:** Simple manipulations on (4.2.9) give the error term, assuming existence of the inverse term under persistence of excitation, as:

$$\tilde{\bar{A}}_{k|k,M} = \bar{A}_{k|k,M} - A = E \left[ k^{-1} \sum_{t=1}^{k} V_t X_{t-1} | \mathcal{Y}_k, M \right] \cdot E \left[ k^{-1} O_k | \mathcal{Y}_k, M \right]_{\text{diag}}^{-1}. \quad (4.4.5)$$

where $V_k = X_k - AX_{k-1}$. Now following the first derivation in the proof of Lemma 4.2.1, we have for some constant $K$

$$E \left[ \left( \sum_{j=1}^{k} V_j X_{j-1} \right) \left( \sum_{t=1}^{k} V_t X_{t-1} \right) \right] \leq K. k$$
Consequently, in addition to $\phi_k = \rho(k)^{-\frac{1}{2}} \sum_{l=1}^{k} V_l X_{l-1}^i$ satisfying $\lim_{k \to \infty} \phi_k = 0$ a.s. for $M$, see (4.2.19), we have the following derivations paralleling those in (4.2.17)

\[
E[(\phi_{k}^{i})^{2}] = \rho(k)^{-1} \sum_{r=1}^{k} (X_{r-1}^{i})^{2} E[(V_{r-1}^{j})^{2} | \mathcal{G}_{t-1}]
\]

\[
\leq \rho(k)^{-1} \sum_{i=1}^{k} K
\]

\[
\leq \frac{K k}{\rho(k)} = \frac{K}{k^{-1} \rho(k)} \text{ for all } i, j. \quad (4.4.6)
\]

Thus

\[
E[(\phi_{k}^{i})^{2}] < \infty \text{ a.s. } (\text{Actually, } \lim_{k \to \infty} E[(\phi_{k}^{i})^{2}] = 0 \text{ a.s. } ) \text{ for all } i, j. \quad (4.4.7)
\]

This is a uniform integrability condition, which together with the property $\lim_{k \to \infty} \phi_k = 0$ a.s. ensures convergence in conditional mean:

\[
E[\phi_k | \mathcal{Y}_k, M] \to 0 \text{ a.s.} \quad (4.4.8)
\]

The convergence (and rate) result (4.4.4) follows for $\tilde{A}_{k|k, M}$ under the excitation condition (4.4.5).

Similarly, the results for $\tilde{C}_{k|k, M}$ and $\tilde{D}_{k|k, M}$ hold.

The convergence results follow from the techniques used for almost sure and mean square convergence in Lemma 4.2.1.

\[ \square \]

We now proceed to present global convergence results where conditional mean estimates based on model estimates are used.

### 4.4.2 Global Convergence

**Notation:** To make use of the convergence rate results of [29] we introduce a $M$ bounded notation. We say $\epsilon$ is $O_M(k^{-1/2})$ if $\epsilon$ is $M$ bounded in the sense that $\sup_k E \left[ \left. \left( \frac{\epsilon}{k^{-1/2}} \right)^2 \right| \mathcal{G} \right] = M_k(q) < \infty$ for all $1 \leq q < \infty$.

**Theorem 4.4.2** Consider the HMM of (4.2.2) (4.2.3) and a particular observation sequence and state sequence outcome, $\{y_k\}$ and $\{X_k\}$, of the HMM with all states in $\{X_k\}$ persistently exciting in that (4.2.11) holds. Consider an sequence of estimated models $\hat{M}_k$, adaptively
update by previous parameter estimates so that \( \widehat{M}_{k+1} = M(\widehat{A}_{k|k}, \widehat{C}_{k|k}, \widehat{D}_{k|k}, \pi_0) \).

We suppose that \( \widehat{M}_k \) is persistently exciting, along with \( M \), in that (4.4.3) holds. Then

\[
\lim_{k \to \infty} \widehat{A}_{k|k}, \widehat{C}_{k|k}, \widehat{D}_{k|k} = A, C, D, \text{ a.s.} \tag{4.4.9}
\]

The errors \( \widehat{A}_{k|k}, \widehat{C}_{k|k}, \widehat{D}_{k|k} \) and \( \widehat{D}_{k|k} \) are \( O_M(k^{-1/2}) \).

**Proof:** The following proof requires that the Markov chain to be asymptotically ergodic. The persistence of excitation condition (4.2.11) implies that the chain is irreducible. Irreducibility together with the results of Seneta [69, Page 128] prove that the chain is asymptotically ergodic. We also require that filters for the conditional mean estimate \( \widehat{J}_{k|k} \), \( \widehat{O}_{k|k} \), and \( \widehat{J}_{k|k} \) are exponentially forgetting which follows by appealing Leroux [46].

First we consider estimation of \( A \) only. Simple manipulations of (4.4.1) give recursions for the parameter estimates \( \widehat{A}_{k|k|k} \), as follows:

\[
\widehat{A}_k = \widehat{A}_{k-1} + \frac{1}{k} (\Delta \widehat{J}^t_k - \widehat{A}_{k-1} \Delta \widehat{O}_k) \left( \frac{1}{k} \widehat{O}_{k|k} \right)^{-1}, \tag{4.4.10}
\]

where we denote \( \widehat{A}_{k|k} \) by the shorthand notation \( \widehat{A}_k \) and define \( \Delta \widehat{J}^t_k := \widehat{J}^t_k - \widehat{J}^t_{k-1} \), \( \Delta \widehat{O}_k := \widehat{O}_{k} - \widehat{O}_{k-1} \), and \( \Delta \widehat{O}_k = \widehat{O}_{k|k} - \widehat{O}_{k-1|k-1} \). Note that \( \widehat{M}_k \) is a function of \( \widehat{A}_{k|k} \), and previous parameter estimates but to simplify the notation we will not always explicitly write this dependence. In particular, when we use the shorthand notation \( \Delta \widehat{J}^t_k \) it will be understood that there is a dependence on \( \widehat{A}_{k|k} \) and previous model estimates.

Convergence of recursion (4.4.10) can be shown by considering an associated ordinary differential equation (ODE) using the approach introduced by Ljung [50, 51] with further contributions made by Kushner [42] and Gerencsér [29]. That is, consider the following ODE,

\[
\frac{d \bar{A}^\tau(\tau, k)}{d\tau} = R^{-1}(\tau, k) f(\bar{A}(\tau, k), k), \quad R(0, k) \geq \delta I, \tag{4.4.11}
\]

where \( k \) is here a fixed parameter and \( \bar{A}^\tau(\tau, k) := \text{col vec}(\bar{A}(\tau, k)) \), where for an arbitrary matrix \( A \) we define \( \text{col vec}(A) := [a_{11}, \ldots, a_{1N}, a_{22}, \ldots, a_{2N}, \ldots, a_{NN}]' \). With \( \bar{A}(\tau, k) \) abbreviated as \( \bar{A} \) we define \( f(\bar{A}, k) \) and \( G(\bar{A}, k) \) as follows

\[
f(\bar{A}, k) := \text{col vec}(E[\Delta \widehat{J}^t_k|A] - \bar{A} \Delta \widehat{O}_k|A]) \quad \text{and} \quad G(\bar{A}, k) := I_N \otimes E[\Delta \widehat{O}_k|A]. \tag{4.4.12}
\]
where $I_N$ is the identity matrix of size $N$ and $\otimes$ is the Kronecker product. Here we have explicitly shown that $\Delta \tilde{\mathbf{J}}_k$ and the other conditional mean estimates depend on the model estimates $\tilde{A}$. Notice that $G$ is block diagonal and the matrix $\tilde{O}_k$ is diagonal.

Consider the function

$$W(\tilde{A}, k) = E \left[ E \left[ \sum_{t=1}^{k} \|X_t - \tilde{A}X_{t-1}\|^2 \left| \tilde{A}, Y_k \right\right] - E \left[ \sum_{t=1}^{k-1} \|X_t - \tilde{A}X_{t-1}\|^2 \left| \tilde{A}, Y_{k-1} \right\right] \right].$$

(4.4.13)

It follows from classical expectation results, including the fact that $E[X|A_1] = E[X|A_1]$ when $A_1 \subset A_2$, that $W(\tilde{A}, k) = E[\|X_k - \tilde{A}X_{k-1}\|^2] \geq 0$. Under asymptotic ergodicity and certain smoothness conditions, that for $k$ sufficiently large the differentiation w.r.t. $\tilde{A}$ and the expectation operations can be interchanged. Hence, for large $k$,

$$\frac{dW(\tilde{A}(\tau, k), k))}{d\tilde{A}(\tau, k)} = -f(\tilde{A}(\tau, k), k).$$

(4.4.14)

It then follows that for large $k$,

$$\frac{dW(\tilde{A}(\tau, k), k)}{d\tau} = \frac{dW(\tilde{A}(\tau, k), k)}{d\tilde{A}(\tau, k)} \frac{d\tilde{A}(\tau, k)}{d\tau} = -f(\tilde{A}(\tau, k), k)R(\tau, k)^{-1}f(\tilde{A}(\tau, k), k).$$

(4.4.15)

Hence, in the limit $k \to \infty$, $\frac{dW(\tilde{A}(\tau, k))}{d\tau} < 0$ because $R(\tau, k)$ is positive definite for all $\tau, k$. Hence for sufficiently large $k$, $W(\tilde{A}, k)$ is a Lyapunov function. Consequently, from Lynamov’s direct method and (4.4.14), $\tilde{A}(\tau, k)$ converges to the set $\{\tilde{A} \lim_{k \to \infty} f(\tilde{A}, k) = 0\}$ a.s.

Under asymptotic ergodicity, in the limit as $k \to \infty$, the set $\{\tilde{A} \lim_{k \to \infty} f(\tilde{A}, k) = 0\}$ is equivalent to the set $\{\tilde{A} \lim_{k \to \infty} E \left[ \hat{f}_{k|k, \tilde{X}_k}(\tilde{A}) \right] - \tilde{A} \tilde{O}_{k|k, \tilde{X}_k}(\tilde{A}) = 0\}$. It follows from Theorem 4.1 that the true model $A$ is a member of this set. Global convergence follows because there is a unique $a_{ij}$ (if there is one) that satisfies the expression $\frac{dW(\tilde{A}(\tau, k))}{d\tilde{a}_{ij}} = E[X_i^\tau X_{i-1}^\tau - a_{ij}X_{i-1}^\tau X_{i-1}^\tau] = 0$. The persistence of excitation conditions imply the existence of a solution, hence $\tilde{A}$ converges uniquely to the true $A$. As a consequence of the asymptotic ergodicity conditions, the ODE convergence result implies convergence of the stochastic difference equation, hence in the limit as $k \to \infty, \tilde{A}_{k|k, \tilde{X}_k}$ converges to the true $A$.

To establish convergence rate we appeal to the results of Gerencsér [29]. In particular we appeal to Theorem 4.1 of [29] which states that under particular conditions the convergence rate of the solution of a differential equation to a fixed point can be $M$-bounded. To apply the results of Gerencsér to the ODE (4.4.11) we note (in his notation) the following constructions:
\[ x = \text{col vec}(\bar{A}), \phi = \left( \text{col vec}(\Delta \hat{J}_{k|\bar{A}}), \text{col vec}(\Delta \hat{O}_{k|\bar{A}}) \right), \]
\[ Q(\phi) = \text{col vec} \left( (\Delta \hat{J}_{k|\bar{A}} - \bar{A} \Delta \hat{O}_{k|\bar{A}}) (\Delta \hat{O}_{k|\bar{A}})^{-1} \right), \quad D_0 = \{ \Delta \hat{J}_{k|\bar{A}}, \Delta \hat{O}_{k|\bar{A}} : \bar{A} \in M_{st} \}, \]
where there is some \( D_{00} \subset D_0 \) that is a compact domain invariant for (4.4.11), and \( M_{st} \) denotes the set of stochastic matrices. Further, as a consequence of our persistence of excitation conditions it can be shown that the conditions of Theorem 4.1 of [29] are satisfied. Hence the Gerencsér results are directly applicable, that is \( \hat{A}_{k|\bar{A},\bar{X}_{tk}} = O_M(k^{-1/2}) \).

We prove convergence when simultaneously estimating \( A, C \) and \( D \) in a similar way. We proceed by defining a new quantity \( \hat{H}_k := \hat{D}^2_{k|\bar{A},\bar{X}_{tk}} + \hat{C}_{k|\bar{A},\bar{X}_{tk}}^2 \) where the squaring is element-wise. Convergence of \( \hat{H}_k \) and \( \hat{C}_{k|\bar{A},\bar{X}_{tk}} \) together will imply convergence of \( \hat{D}_{k|\bar{A},\bar{X}_{tk}} \). Simple manipulations of (4.2.14) and (4.2.15) gives recursions for the \( C \) and \( H \) estimates as
\[
\begin{align*}
\hat{C}_k &= \hat{C}_{k-1} + \frac{1}{k} (\Delta \hat{T}_k(y) - \hat{C}_{k-1} \Delta \hat{O}_k) \left( \frac{1}{k} \hat{O}_{k|\bar{A},\bar{X}_{tk}} \right)^{-1} \\
\hat{H}_k &= \hat{H}_{k-1} + \frac{1}{k} (\Delta \hat{T}_k(y)^2 - \hat{H}_{k-1} \Delta \hat{O}_k) \left( \frac{1}{k} \hat{O}_{k|\bar{A},\bar{X}_{tk}} \right)^{-1}
\end{align*}
\]
(4.4.16)

where \( \hat{C}_k \) is shorthand notation for \( \hat{C}_{k|\bar{A},\bar{X}_{tk}} \) and \( \Delta \hat{T}_k(f) := \hat{T}_{k|\bar{A},\bar{X}_{tk}}(f) - \hat{T}_{k-1|\bar{A},\bar{X}_{tk-1}}(f) \).

Now consider an ODE associated with the \( A, C \) and \( H \) recursions
\[
\begin{align*}
\frac{d\bar{\theta}(\tau, k)}{d\tau} &= R^{-1}(\tau, k) f(\bar{\theta}(\tau, k), k) \\
\frac{dR(\tau, k)}{d\tau} &= G(\bar{\theta}(\tau, k), k), \quad R(0, k) \geq \delta I,
\end{align*}
\]
(4.4.17)

where here \( k \) is fixed and with \( f(\bar{\theta}(\tau, k), k) \) and \( G(\bar{\theta}(\tau, k), k) \) defined as follows
\[
\begin{align*}
f(\bar{\theta}(\tau, k), k) &:= \left[ \begin{array}{c} \text{col vec}(E[\Delta \hat{T}_k(y) - \hat{A}(\tau, k) \Delta \hat{O}_k]) \\ \text{col vec}(E[\Delta \hat{T}_k(y)^2 - \hat{C}(\tau, k) \Delta \hat{O}_k]) \\ \text{col vec}(E[\Delta \hat{T}_k(y)^2 - \hat{H}(\tau, k) \Delta \hat{O}_k]) \end{array} \right] \\
G(\bar{\theta}(\tau, k), k) &:= I_{N+2} \otimes E[\Delta \hat{O}_k] \quad \quad \quad (4.4.18)
\end{align*}
\]

where
\[
\bar{\theta}(\tau, k) = \left[ \begin{array}{c} \text{col vec}(\bar{A}(\tau, k)) \\ \text{col vec}(\bar{C}(\tau, k)) \\ \text{col vec}(\bar{H}(\tau, k)) \end{array} \right]
\]

With \( \bar{\theta}(\tau, k) \) abbreviated as \( \bar{\theta} \), consider the Lyapunov function,
\[
\bar{W}(\bar{\theta}, k) = E \left[ E \left[ \sum_{t=1}^{k} \left| X_t - \bar{A} X_{t-1} \right|^2 \left| \bar{A}, \gamma_k \right] \right] - E \left[ \sum_{t=1}^{k-1} \left| X_t - \bar{A} X_{t-1} \right|^2 \left| \bar{A}, \gamma_{k-1} \right] \right] + E \left[ \sum_{t=1}^{k} \sum_{i=1}^{N} (y_t - \bar{C}_i X_t)^2 \left| \bar{C}, \gamma_k \right] \right] - E \left[ \sum_{t=1}^{k} \sum_{i=1}^{N} (y_t - \bar{C}_i X_t)^2 \left| \bar{C}, \gamma_{k-1} \right] \right]
\]
\[ + E \left[ \sum_{t=1}^{k} \sum_{i=1}^{N} (y_t^2 - \hat{H}_t x_t^2) \left\| \hat{H}_t, Y_k \right\| \right] - E \left[ \sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_t^2 - \hat{H}_t x_t^2) \left\| \hat{H}_t, Y_{k-1} \right\| \right] \]

(4.4.19)

In a similar manner to the above case for estimation of \( A \), it follows that \( \bar{W}(\tilde{\theta}(\tau, k), k) \geq 0 \) for sufficiently large \( k \) and that \( \frac{d \bar{W}(\tilde{\theta}(\tau, k), k)}{d\tilde{\theta}(\tau, k)} = -f(\tilde{\theta}(\tau, k), k) \), hence \( \frac{d \bar{W}(\tilde{\theta}(\tau, k), k)}{d\tau} < 0 \) for sufficient large \( k \). Hence, \( \bar{W}(\tilde{\theta}, k) \) is a Lyapunov function for sufficiently large \( k \).

By straightforward application of Lyapunov’s direct method it follows that the estimates of \( A, C \) and \( D \) converge to the set \( \{ \hat{A}, \hat{C}, \hat{H} \} \) \( \lim_{k \to \infty} \frac{d \bar{W}(\tilde{\theta}(\tau, k), k)}{d\tau} = 0 \). The set \( \{ \hat{A}, \hat{C}, \hat{H} \} \) \( \lim_{k \to \infty} \frac{d \bar{W}(\tilde{\theta}(\tau, k), k)}{d\tau} = 0 \) is equivalent to the sets \( \{ \hat{A} \lim_{k \to \infty} E \left[ \tilde{J}_{k|k, \hat{X}_{ka}} - \hat{A} \hat{O}_{k|k, \hat{X}_{ka}} \right] = 0 \} \),
\( \{ \hat{C} \lim_{k \to \infty} E \left[ \tilde{T}_{k|k, \hat{X}_{ka}}(y) - \hat{A} \hat{O}_{k|k, \hat{X}_{ka}} \right] = 0 \} \) and \( \{ \hat{H} \lim_{k \to \infty} E \left[ \tilde{T}_{k|k, \hat{X}_{ka}}(y^2) - \hat{R} \hat{O}_{k|k, \hat{X}_{ka}} \right] = 0 \} \).

It follows from Theorem 4.4.1 that the true model parameters \( A, C \) and \( D \) are members of these convergence sets, remembering that \( \tilde{D}_{k|k, \hat{X}_{ka}} = \sqrt{H_k - C_k^2} \). Global convergence of \( \tilde{A}_{k|k, \hat{X}_{ka}}, \tilde{C}_{k|k, \hat{X}_{ka}} \) and \( \tilde{D}_{k|k, \hat{X}_{ka}} \) follows. The theorem convergence rates follow with appropriate choices of \( \phi, Q(\phi), D_0 \) and \( D_00 \) in a similar way to above by appealing to [29].

\( \square \)

The form of the Lyapunov function \( \bar{W}(\tilde{\theta}, k) \) has a significant interpretation in terms of results of previous chapters. In particular, we note that the cost (2.4.10) on Page 35 and the cost (3.3.7) on Page 56 are equivalent to two of the three terms in Lyapunov function (4.4.19). It is surprising that these cost functions reappear in the Lyapunov function (4.4.19) and hence two alternative approaches are so strongly connected. The form of Lyapunov function (4.4.19) suggests that the cost functions used in Chapters 2 and 3 are the appropriate cost functions for identification of HMMs. The connection between the RPE algorithms of Chapters 2 and 3 and the on-line algorithm of this chapter is analogous to the connection between the Healy-Westmacott and EM algorithm.

**Remarks**

1. It is clear that the correct motivation for the cost functions used in the previous chapters is from an incomplete data or missing data problem perspective. Viewing the identification problem as a minimization of the usual least squares (prediction error) cost does not lead to correct identification in this case, hence the problems of the approach taken in [13] which are discussed in [16].
2. The use of the $\sum_{t=1}^{k} \sum_{l=1}^{N} (y_t - C^i X_l)^2$ term rather than a $\sum_{t=1}^{k} (y_t - C X_t)^2$ term improves the numerical conditioning of the problem as the noise level decreases, see Chapter 2

4.4.3 Reduced Complexity Algorithms

One way to reduce the computational requirements of the overall estimation algorithm (4.4.1) is by reducing the computational effort used to produce conditional mean estimates. There is a subset of models $\{\tilde{M}\}$ for which calculation of the conditional mean estimates is computationally easier than for other models. Hence, the key idea of this section is to calculate the conditional mean estimates of $J_k$ and $O_k$ corresponding to one of these special models $\tilde{M}$ (or a sequence of such models) rather than generic model “estimates” such as the adaptive $M_k$ above. This will reduce the computational effort required to implement the overall estimation algorithm (4.4.1).

An example of one such $\tilde{M}$ is the i.i.d. sequence model which is a subset of all valid HMMs. For the purpose of generating the conditional mean estimates, $\tilde{F}_{k|k,\bar{X}_{n-1}}$ and $\tilde{O}_{k|k,\bar{X}_{n-1}}$, the state sequence can be modelled as an i.i.d. sequence, leading to a model estimate with $\tilde{M}_k = M_{iid}$ for all $k$ as

$$M_{iid} = \{A_{iid}, C, D, \pi_0\}, \quad A_{iid} = \frac{1}{N} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}. \quad (4.4.20)$$

The filters now denoted $\tilde{F}_{k|k,M_{iid}}$ and $\tilde{O}_{k|k,M_{iid}}$ require $O(N^3)$ calculations per time instant. This is a reduction from the $O(N^4)$ calculations per time instant required to implement the general filters $\tilde{F}_{k|k,\bar{X}_{n-1}}$ and $\tilde{O}_{k|k,\bar{X}_{n-1}}$ using estimates $\bar{A}_{k|k,\bar{X}_{n-1}}$.

The filters for the model set $M_{\alpha\beta} = \{\alpha A_{iid} + \beta I_{N\times N}, C, D, \pi_0\}$ for scalar $\alpha, \beta$ where $\{\alpha, \beta : 1^T A = 1\}$ also requires only $O(N^3)$ calculations per time instant. This set can approximate a larger class of models and results in reduced estimation error.

Remark

1. Even though $A = I$ may appear a likely candidate model of $\{\tilde{M}\}$ for reducing complexity the persistently excitation condition (4.4.3) is not satisfied for this model and the convergence no longer holds.
2. Correct estimation occurs in low noise because \( \hat{X}_{k|k,\hat{\mathcal{M}}_k} \approx X_k \).

Another Suboptimal Simplifications

Let us denote suboptimal estimates of \( J_k, O_k, T_k \) as follows

\[
J_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} = \sum_{t=1}^{k} \mathbb{E} \left[ X_{t-1} X_t^t \mid \mathcal{Y}_t, \hat{\mathcal{M}}_t \right]
\]

(4.4.21)

\[
O_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} = \sum_{t=1}^{k} \left( \mathbb{E} \left[ X_{t-1} \mid \mathcal{Y}_t, \hat{\mathcal{M}}_t \right] \right)_{\text{diag}}
\]

(4.4.22)

\[
T_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} (f) = \sum_{t=1}^{k} f(\mathcal{Y}_t) \mathbb{E} \left[ X_{t-1} \mid \mathcal{Y}_t, \hat{\mathcal{M}}_t \right].
\]

(4.4.23)

Then the corresponding suboptimal model estimates are

\[
A_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} = J_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} \left( O_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} \right)^{-1}
\]

(4.4.24)

\[
C_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} = T_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} (g) \left( O_{k|k,\hat{\mathcal{M}}_k}^{\text{sub}} \right)^{-1}
\]

(4.4.25)

and likewise for \( D_k^{\text{sub}} \), via (4.2.16).

These suboptimal estimates require \( O(N^2) \) calculations per time instant to calculate.

4.5 Simulations Studies

In this section, simulation studies are presented for a selection of the adaptive HMM schemes.

4.5.1 Conditional Mean Estimates

Low Noise A 10000 point, 2-state Markov process and observation process of the form (4.2.2) and (4.2.3) is generated with parameters: \( a_{ii} = 0.8 \), \( i = 1, 2 \) and \( a_{ij} = 0.2 \) for all \( i, j \neq j \), \( C = [2, 4]^t \) and \( D = [0.1, 0.1]^t \) (that is state independent noise of variance 0.01).

Adaptive estimation of \( A, C \) and \( D \) is performed using (4.4.1). The model estimate \( \hat{\mathcal{M}}_k \) is adaptively updated at each time instant. Note that the recursions for \( \hat{X}_{k|k,\hat{\mathcal{M}}_k}, \hat{\mathcal{F}}_{k|k,\hat{\mathcal{M}}_k} \) and \( \hat{\mathcal{O}}_{k|k,\hat{\mathcal{M}}_k} \) given in [18] are used and that no backwards step is required.

Initial estimates of the HMM parameter are: \( a_{ii} = 0.1 \), \( C = [1.5, 3]^t \) and \( D = [\sqrt{0.05}, \sqrt{0.05}] \).

The final estimates for \( A \) and \( C \) are; \( \bar{A} = [0.7924, 0.2076; 0.2038, 0.7962] \) and \( \bar{C} = [1.9971, 3.9990]^t \).
The estimates of all parameter converge to the correct value from the initial estimates. To illustrate the global convergence properties, a variety of initial estimates are tried. Convergence occurred from all initial estimates tried.

**Medium Noise**  The same HMM model is generated except with $D = [\sqrt{0.1}, \sqrt{0.1}]'$. Initial estimates of the HMM parameter are: $a_{ii} = 0.1$, $C = [1, 5]'$ and $D = [\sqrt{0.5}, \sqrt{0.5}]'$. Adaptive estimation of $A$, $C$, and $D$ is performed using (4.4.1). The model estimate $\hat{M}_k$ is adaptively updated at each time instant.

Figures (4.1) and (4.2) show the evolution of the estimation error for several key parameters. To illustrate the global convergence properties a variety of initial estimates are been tried. Convergence occurred from all initial estimates tried.

The final estimated value for the noise variance is 0.0171 (real value=0.1).

### 4.5.2 Suboptimal Adaptive Schemes

For the HMMs above, the suboptimal schemes using jump, occupation and transition estimates (4.4.21), (4.4.22) and (4.4.23) respectively and the estimation equations (4.4.24) and (4.4.25) are used to obtain adaptive estimates of $A$ and $C$.

**Low Noise**  The same HMM model is generated again except with $D = [\sqrt{0.01}, \sqrt{0.01}]'$. Initial estimates of the HMM parameter are: $a_{ii} = 0.1$, $C = [1.5, 3]'$. Adaptive estimation of $A$, $C$ is performed using (4.4.1). The model estimate $\hat{M}_k$ is adaptively updated at each time instant.

The final estimated values for $A$ and $C$ are, $\hat{A} = [0.7938, 0.2062; 0.2053, 0.7949]$ and $\hat{C} = [1.9981, 3.9997]'$.

**Medium Noise**  The same HMM model is generated again except with $D = [\sqrt{0.1}, \sqrt{0.1}]'$. Initial estimates of the HMM parameter are: $a_{ii} = 0.1$, $C = [1, 5]'$. Adaptive estimation of $A$, $C$ was performed using (4.4.1). The model estimate $\hat{M}_k$ is adaptively updated at each time instant.

Figures 4.3 and 4.4 show the evolution of the estimation error for several key parameters. To illustrate the global convergence properties a variety of initial estimates are tried. Convergence occurred from all initial estimates tried.
Comparison with optimal filtering The optimal and suboptimal adaptive estimation schemes are compared on the same HMM. A medium noise HMM is generated with the parameters used previously and the all the parameters are estimated on-line. Figure 4.5 shows a comparison of estimation errors resulting from estimation of $a_{11}$. The performance of the two adaptive schemes is similar.

4.5.3 Other Markov Chains

Low Inertia HMM To further test the adaptive schemes consider a 1000 point, 2-state HMM with low inertia parameter values $a_{ii} = 0.2$, $i = 1, 2$ and $a_{ij} = .8$ for all $i, j$ $i \neq j$, $C = [2, 4]$ and $D = [0.1, 0.1]'$.

This HMM is generated and adaptive estimation of $A, C$ is performed using (4.4.1).

Figure 4.6 shows convergence of the adaptive estimate of $a_{11}$. All other parameter estimates converge similarly.

Mixed Inertia HMM Consider a more general 2-state HMM with parameter values $a_{11} = 0.8$, $a_{21} = 0.2$, $a_{12} = 0.7$, $a_{22} = 0.3$, $C = [2, 4]$ and $D = [0.1, 0.1]'$.

This HMM is generated and adaptive estimation of $A, C$ is performed using (4.4.1).

After 10000 points convergence has occurred, the estimation error in the $A$ estimates is $1 \times 10^{-3} [0.9215, -0.0370, -0.9215, 0.0370]$, and in $C$ estimates is $[-0.00482, 0.00305]$.

4.5.4 A Comparison with Other Methods

Our experience is that the new adaptive schemes work at least as well as earlier schemes, such as those in [39, 13] and Chapters 2 and 3, and often better. However, sample comparisons are not included here because they could be misleading and are perhaps dependent on initialization and various tuning parameters.

4.5.5 Reduced Complexity Estimation

A 3000 point, 2-state HMM was generated with parameters: $a_{ii} = 0.8$, $i = 1, 2$ and $a_{ij} = 0.2$ $\forall i, j \neq j$, $C = [2, 4]'$ and $D = [0.1, 0.1]'$. Estimation was performed in two ways: using the simplified model approximation (4.4.20), and using $\tilde{M}_k$ corresponding to the best available parameter estimates. At this noise level little bias is introduced by these approximation.
4.5.6 Threshold

To examine the bias introduced into the parameter estimation a 10000 point, 2-state HMM generated with parameters: \( a_{ii} = 0.8, \ i = 1,2 \) and \( a_{ij} = 0.2 \ \forall i,j \ i \neq j \), \( C = [2,4] \)' was simulated at various SNRs. Estimation of \( A \) is performed assuming \( M = M_{\text{id}} \). Figure 4.7 shows estimation error versus SNR. From this figure it appears that for \( \text{SNR} > 12 \) or \( D < 0.2 \) the bias introduced is insignificant.

4.5.7 More Complicated Approximations

A 10000 point, 2-state HMM was generated with parameters: \( a_{ii} = 0.8, \ i = 1,2 \) and \( a_{ij} = 0.2 \ \forall i,j \ i \neq j \), \( C = [2,4] \)' and \( D = 0.7 \) or \( \text{SNR}=6.3 \). Estimation of \( A \) was performed using three methods: using the model set \( M_{\alpha \beta} \), using the model set \( M_{\text{id}} \), and using the adaptive scheme (4.4.1) Figure 4.8 shows the convergence of parameters estimates over time. At this noise level it appears using models \( M_{\alpha \beta} \) reduces the estimation error.

4.5.8 High Noise and High Order Models

Current research is underway to clarify the relative power of the new algorithms in more extreme situations.

4.6 Conclusions

Global convergence results have been developed for a new adaptive HMM parameter estimation scheme. The online adaptive algorithm appears related to the off-line Baum-Welch EM algorithm; however, convergence results are developed using standard martingale properties, martingale convergence results, the Kronecker lemma and an ordinary differential equation approach. We emphasize that under reasonable excitation conditions, there is consistent estimation when model estimates are updated adaptively. Almost sure rates of convergence also are guaranteed under the persistence of excitation conditions. Simulation studies are presented that highlight convergence properties of the adaptive algorithm.
Figure 4.1: Estimation error in $\hat{a}_{11}$ in medium noise

Figure 4.2: Estimation error in $\hat{C}_1$ in medium noise
Figure 4.3: Suboptimal Estimation of $\widehat{\mathcal{C}}_1$ in medium noise

Figure 4.4: Suboptimal Estimation of $\widehat{\mathcal{C}}_1$ in medium noise
Figure 4.5: Comparison of optimal and suboptimal filters

Figure 4.6: Estimation for unusual HMM
Figure 4.7: The bias after 10000 points.

Figure 4.8: An empirical comparison
Chapter 5

Estimation of Linear and Hybrid Linear Systems

5.1 Introduction

There are many linear system identification methods [31, 41, 52, 34, 48, 3, 72]. In fully observed linear systems, the linear relationship between input and output signals ensures efficient identification and readily established global convergence of parameter estimates, see [34, 48]. In partially observed systems, similar convergence results are more difficult to obtain, and there is scope for further research.

Additionally, the class of systems modelled as stochastic dynamical systems with time-varying parameters is important for modelling some complex physical systems [77]. Abrupt changes in system behaviour can occur when a system changes from one mode to another. In such a situation, the state-space has a natural decomposition as a hybrid combination of continuous-valued states and discrete-valued states. In other applications, it may be possible to approximate a non-linear system as a linear system which switches between different specific dynamics. This type of hybrid linear model has been used in applications such as systems subject to failure [57], and communication systems [18].

Recently, almost surely convergent parameter estimation algorithms have been proposed for hidden Markov models in discrete and continuous time, Chapter 4 and [23], and linear systems in continuous time [24]. In particular, the almost surely convergent parameter estimators proposed for hidden Markov models in discrete time are based on estimates for functions of the Markov state [20]. These estimators and their convergence proofs suggest a similar approach for partially observed linear stochastic systems.
While, exact finite dimensional filters in discrete-time have been proposed for a class of hybrid systems known as Markov jump linear systems [19, 37, 38], however, the dimension of these filters increases with time. Markov jump linear systems (or what are called hybrid linear systems in this thesis) are linear systems whose parameters change abruptly according to the evolution of a Markov chain.

In this chapter, standard martingale convergence results, the Kronecker lemma and ordinary differential equation (ODE) methods (see [50, 51, 42, 29]) are used to study recursive parameter estimation for partially-observed, discrete-time linear stochastic systems and for linear systems modulated by a Markov chain. Preliminary almost sure convergence results and convergence rates are established for both linear and hybrid linear systems using martingale convergence results and the Kronecker lemma when the state is known or conditional mean estimates are available. Complete global convergence results are established for partially observed linear systems using an ODE approach.

The chapter is organized as follows: in Section 5.2, the partially observed discrete-time linear system model is introduced and new recursive parameter estimators are proposed for the special case when the state is measured directly. Almost sure convergence of these parameter estimators is established via martingale convergence results and the Kronecker lemma. In Section 5.3, parameter estimation is discussed when the state is not measured directly. We develop the relevant conditional mean filters, propose parameter estimators and establish almost sure convergence results using an ODE approach. In Section 5.4, parameter estimators for hybrid linear systems are presented. In Section 5.5, some simulation studies are presented. Conclusions are given in Section 5.6.

5.2 Dynamics

Consider a probability space \((\Sigma, \mathcal{F}, P)\); suppose \(\{x_\ell\}, \ell \in \mathbb{Z}^+\) is a discrete-time linear stochastic process, taking values in \(\mathbb{R}^N\), with dynamics given by

\[
x_{k+1} = Ax_k + Bw_{k+1}, \quad x_0 \in \mathbb{R}^{N \times 1}.
\]

(5.2.1)

Here \(k \in \mathbb{Z}^+, \ A \in \mathbb{R}^{N \times N}, B \in \mathbb{R}^{N \times 1}\) and \(\{w_\ell\}, \ell \in \mathbb{Z}^+,\) is a sequence of independent and identically distributed \(N(0,1)\) scalar random variables. The probability density of the noise, \(B w_\ell\), is denoted by \(\psi\).
The state process \( x_\ell, \ell \in \mathbb{Z}^+ \), is observed indirectly via the scalar observation process \( \{ y_\ell \}, \ell \in \mathbb{Z}^+ \), given by

\[
y_\ell = C x_\ell + D v_\ell.
\]  
(5.2.2)

Here for \( k \in \mathbb{Z}^+ \), \( C \in \mathbb{R}^{1 \times N} \), \( D \in \mathbb{R} \) and \( \{ v_\ell \}, \ell \in \mathbb{Z}^+ \), is a sequence of independent and identically distributed \( \mathcal{N}(0,1) \) scalar random variables. The probability density of the noise, \( Dv_\ell \), is denoted by \( \psi \). We assume that \( x_0 \), \( \{ v_\ell \} \) and \( \{ w_\ell \} \) are mutually independent.

Let \( \mathcal{F}_k = \sigma\{x_0, x_1, \ldots, x_k\} \) and \( \mathcal{Y}_k = \sigma\{y_0, y_1, \ldots, y_k\} \) denote the filtrations generated by \( x \) and \( y \) respectively. Also, let \( \mathcal{G}_k = \sigma\{x_0, \ldots, x_k, y_0, \ldots, y_k\} \) denote the filtration generated by \( x \) and \( y \).

The model described by (5.2.1), (5.2.2) is denoted by

\[
\lambda = \lambda(A, B, C, D, x_0). 
\]  
(5.2.3)

### 5.2.1 Measure Change

Define the probability measure \( \bar{P} \) such that the \( \mathcal{G}_k \) restriction of the Rayon-Nikodym derivative of \( P \) with respect to \( \bar{P} \) is

\[
\frac{dP}{d\bar{P}} \bigg|_{\mathcal{G}_k} = \Gamma_k = \prod_{\ell=1}^k \gamma_\ell, \quad \text{where} \quad \gamma_\ell = \frac{\psi(Bw_\ell)\phi(Dv_\ell)}{\psi(x_\ell)\phi(y_\ell)}. 
\]  
(5.2.4)

**Theorem 5.2.1** Under the measure \( \bar{P} \):

1. \( \{x_\ell\} \) and \( \{y_\ell\} \) are independent random variables with densities \( \psi \) and \( \phi \), respectively. \( \{x_\ell\} \) and \( \{y_\ell\} \) are also independent of each other.
2. If \( f_\ell \) is a \( \mathcal{G} \) adapted sequence then the following abstract version of Bayes theorem holds:

\[
E[f_\ell|\mathcal{Y}_k] = \frac{E[\Gamma_\ell f_\ell|\mathcal{Y}_k]}{E[\Gamma_\ell|\mathcal{Y}_k]} 
\]  
(5.2.5)

where \( E[.] \) denotes expectation in \( \bar{P} \) and \( \Gamma_\ell \) denotes the reverse measure change. See [18] for information about the reverse measure change.

**Proof:** see [18].

Identity (5.2.5) enables us to obtain the conditional expectation \( E[f_\ell|\mathcal{Y}_k] \) if we know the unnormalized conditional expectation \( E[\Gamma_\ell f_\ell|\mathcal{Y}_k] \).
\section*{5.2.2 Parameter Estimation - Full Observations}

In this subsection we assume that both \( \{x_k\} \) and \( \{y_k\} \) are fully observed. The results in this section for the full observation case are presented as a stepping stone to the more interesting and general results of Section 5.3.

From (5.2.1), by post-multiplication by \( x_i' \) and summing, we obtain

\[
\sum_{i=1}^{k} x_i x_i' = A \sum_{i=1}^{k} x_{i-1} x_i' + B \sum_{i=1}^{k} w_i x_i'.
\]  
(5.2.6)

Here the prime \( ' \) denotes the transpose operation. Now consider the matrices

\[
J_k = \sum_{i=1}^{k} x_i x_i' , \quad O_k = \sum_{i=1}^{k} x_{i-1} x_i' \quad \text{and} \quad M_k = \sum_{i=1}^{k} w_i x_i'.
\]  
(5.2.7)

From (5.2.6) we see that \( J_k = AO_k + BM_k \). Assuming that \( x_k \) is measurable, a reasonable estimate for \( A \) is therefore

\[
\hat{A}_k = J_k O_k^{-1}
\]  
(5.2.8)

whenever \( O_k^{-1} \) exists. The error in this estimate is \( \hat{A}_k - A = BM_k O_k^{-1} \).

Similarly, from (5.2.2) by post-multiplication by \( x_i' \) and summing we obtain

\[
\sum_{i=1}^{k} y_i x_i' = A \sum_{i=1}^{k} x_{i-1} x_i' + D \sum_{i=1}^{k} v_{i-1} x_i'.
\]  
(5.2.9)

Now consider the vectors

\[
T_k = \sum_{i=1}^{k} y_{i-1} x_i' \quad V_k = \sum_{i=1}^{k} v_{i-1} x_i'.
\]  
(5.2.10)

Then (5.2.9) can be written \( T_k = CO_k + DV_k \). A reasonable estimate for \( C \) is therefore

\[
\hat{C}_k = T_k O_k^{-1}
\]  
(5.2.11)

whenever \( O_k^{-1} \) exists. The error in this estimate is \( \hat{C}_k - C = DV_k O_k^{-1} \).

To obtain estimates for \( B \) and \( D \) we square element-wise and sum over \( k \) equations (5.2.1) and (5.2.2) individually giving

\[
\left( (b^1)^2, \ldots, (b^N)^2 \right) \sum_{t=1}^{k} w_t^2 = \sum_{t=1}^{k} \left( x_k^1 - \sum_{i=1}^{N} a_{1i} x_{t-1}^i \right)^2, \ldots, \]
\[
\left( x_k^N - \sum_{i=1}^{N} a_{Ni} x_{t-1}^i \right)^2 \right)
\]  
(5.2.12)
and
\[ D^2 \sum_{t=1}^{k} v_t^2 = \sum_{t=1}^{k} (y_t - \sum_{i=1}^{N} C_i x_i)^2, \]  
(5.2.13)
where \( B = (b^1, \ldots, b^N)' \). From (5.2.12) and (5.2.13) we obtain estimates for \( B \) and \( D \) respectively as
\[
\left( (\hat{b}_k^1)^2, \ldots, (\hat{b}_k^N)^2 \right) = \frac{1}{k} \sum_{t=1}^{k} \left( (x_k^1 - \sum_{i=1}^{N} a_{1i} x_{t-1}^i)^2, \ldots, \right.

\left. (x_k^N - \sum_{i=1}^{N} a_{Ni} x_{t-1}^i)^2 \right)
\]  
(5.2.14)
and
\[
\hat{D}_k^2 = k^{-1} \sum_{t=1}^{k} (y_t - \sum_{i=1}^{N} C_i x_t^i)^2.
\]  
(5.2.15)
In the following section we investigate the convergence of these estimators.

### 5.2.3 Almost Sure Convergence

In this subsection we discuss the convergence of these estimators. Before proceeding to our convergence results we first state a stability condition for a linear system.

A time-invariant system with state transition matrix \( A \) is stable if the following condition holds
\[
\sigma_{\text{max}}(A) < 1,
\]  
(5.2.16)
where \( \sigma_{\text{max}}(A) \) is the largest magnitude of the eigenvalues of the matrix \( A \). Note, that this is a sufficient but not necessary condition for stability.

The system (5.2.1),(5.2.2) is said to be persistently excited if
\[
\lim_{k \to \infty} \sup_{k} \left( \frac{1}{k} O_k \right)^{-1}
\]  
exists.
(5.2.17)

Now consider
\[
R_k := \sum_{i=1}^{k} \rho(i)^{-\frac{1}{\alpha}} \Delta M_i,
\]
where \( \Delta M_i := M_i - M_{i-1} = w_i x_{i-1} \) and here the function \( \rho(k), k \geq 0 \), is positive, non-decreasing and such that \( \lim_{k \to \infty} \sum_{i=0}^{k} \rho(i)^{-1} = \lambda < \infty \). An example of \( \rho(k) \) satisfying this requirement is \( \rho(k) = \max(1, k^{lnk(lnk)^{\alpha}}) \), for any \( \alpha > 1 \).

**Lemma 5.2.1** Suppose the system (5.2.1),(5.2.2) is stable such that (5.2.16) holds. Then \( R_k \) is a vector whose elements, \( R_k^i \) for \( i = 1, \ldots, N \), are square integrable martingales so that \( \lim_{k \to \infty} R_k^i = \lambda^i(\omega) < \infty \) exists a.s.
Proof: $R_k^i$ is a martingale because $E[R_k^i | \mathcal{F}_{k-1}] = E[\rho(k)\Delta M_k^i + R_{k-1}^i | \mathcal{F}_{k-1}] = R_{k-1}^i$, where $\Delta M_k^i$ is the $i$th element of $\Delta M_k$. Also, the $R_k^i$ are bounded in $L^2$ because

$$E \left[ (R_k^i)^2 \right] = E \left[ \sum_{\ell=1}^{k} \rho(\ell)^{-1} (\Delta M_k^i)^2 \right] = E \left[ \sum_{\ell=1}^{k} \rho(\ell)^{-1} E \left[ w_\ell^i (x_{\ell-1}^i)^2 | \mathcal{F}_{\ell-1} \right] \right] = \sum_{\ell=1}^{k} \rho(\ell)^{-1} E \left[ (x_{\ell-1}^i)^2 \right] < \infty.$$

Here $x_{i-1}^i$ is the $i$th element of $x_{\ell-1}$. We have used that $w_\ell$ and $x_{\ell-1}$ are uncorrelated, that $E[w_\ell^i | \mathcal{F}_{\ell-1}] = 1$ and that for stable systems $E[(x_k^i)^2] < B_\infty$ for all $k, i$ for some $B_\infty < \infty$. By standard martingale results, [59, 64],

$$\lim_{k \to \infty} R_k^i = \eta^i(\omega) < \infty \text{ a.s. for } i = 1, \ldots, N.$$

\[ \square \]

**Lemma 5.2.2** Suppose the system (5.2.1), (5.2.2) is stable then

$$\lim_{k \to \infty} \rho(k)^{-\frac{1}{2}} \sum_{i=1}^{k} \Delta M_i = 0_{1 \times N} \text{ a.s.}$$

where $0_{1 \times N}$ is the $(1 \times N)$ vector of all zeros. That is,

$$\lim_{k \to \infty} \rho(k)^{-\frac{1}{2}} M_k = 0_{1 \times N} \text{ a.s.}$$

**Proof:** Follows from Lemma 5.2.1 and by applying the Kronecker Lemma to each element of $R_k$, see also [53, 64].

\[ \square \]

**Theorem 5.2.2** Consider the linear system (5.2.1), (5.2.2) and suppose the $x_k$ and $y_k$ are both observed. Suppose (5.2.16) and (5.2.17) hold. Then

$$\lim_{k \to \infty} \tilde{A}_k, \tilde{B}_k, \tilde{C}_k, \tilde{D}_k = A, B, C, D \text{ a.s. and}$$

the almost sure convergence rate is at least $(k^{-2} \rho(k))^{\frac{1}{2}}$.

**Proof:** First consider the error in estimation of $A$, that is $\tilde{A}_k - A = B(\frac{1}{k} M_k)(\frac{1}{k} O_k)^{-1}$. From Lemma 5.2.2 we have that

$$\lim_{k \to \infty} \frac{1}{k} M_k = 0_{1 \times N} \text{ a.s.} \quad (5.2.20)$$
at a rate of \( \left( \frac{a(k)}{k^2} \right)^{\frac{1}{2}} \). Under the persistently exciting assumption \( \lim_{k \to \infty} \sup_k \left( \frac{1}{k} O_k \right)^{-1} \) exists, so

\[
\lim_{k \to \infty} \tilde{A}_k := \tilde{A}_k - A = 0_{N \times N} \text{ a.s.}.
\]

Indeed \( \tilde{A}_k \to 0_{N \times N} \) a.s. at a rate \( \left( \frac{a(k)}{k^2} \right)^{\frac{1}{2}} \). The result for \( C \) follows similarly using that \( x_{t-1} \) and \( u_{t-1} \) are uncorrelated and that \( E[u_k^2] = 1 \). The convergence results for \( B \) and \( D \) follow because \( E[w_k^2] = 1 \) or equivalently \( \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} w_i^2 = 1 \).

\( \Box \)

**Remarks**

1. Similar convergence results for fully observed stochastic linear systems are common, see [34, 48].

We next propose parameter estimators for partially-observed linear systems based on the conditional mean estimates of \( J_k, O_k \) and \( T_k \). Convergence results are also presented.

### 5.3 Conditional Mean Estimates

In this section we consider parameter estimators based on conditional mean estimates in lieu of the states. Consider again the system (5.2.1), (5.2.2). We define a model set, \( \Lambda \), of allowable model estimates \( \Lambda_k \) and assume that the “correct” model \( \lambda \) lies in the model set \( \Lambda \).

The model set considered in this chapter is \( \Lambda = \{ \lambda(A, B, C, D, x_0) : N^{th} \text{ order model} \} \). Let us denote the associated conditional mean estimates based on the signal generating system \( \lambda \) as in (5.2.3), also termed the “correct” model, and the associated conditional mean estimates based on a possibly time varying estimate \( \tilde{\Lambda}_k = \{ \tilde{\lambda}_1, \ldots, \tilde{\lambda}_k \} \), also termed the “incorrect” model, respectively as

\[
\tilde{J}_{k|\lambda} = E[J_k|Y_k, \lambda], \quad \tilde{O}_{k|\lambda} = E[O_k|Y_k, \lambda], \quad \tilde{T}_{k|\lambda} = E[T_k|Y_k, \lambda], \quad \text{(5.3.1)}
\]

and

\[
\tilde{J}_{k|\tilde{\Lambda}_k} = E[J_k|Y_k, \tilde{\Lambda}_k], \quad \tilde{O}_{k|\tilde{\Lambda}_k} = E[O_k|Y_k, \tilde{\Lambda}_k], \quad \tilde{T}_{k|\tilde{\Lambda}_k} = E[T_k|Y_k, \tilde{\Lambda}_k]. \quad \text{(5.3.2)}
\]

Initially, we consider the (somewhat artificial) special case when it is assumed that the conditional mean estimates based on the correct model are available. We consider the following
parameter estimates:
\[
\hat{A}_{k|k,\lambda} = \hat{J}_{k|k,\lambda} \hat{O}_{k|k,\lambda}^{-1} \quad \text{and} \quad \hat{C}_{k|k,\lambda} = \hat{T}_{k|k,\lambda} \hat{O}_{k|k,\lambda}^{-1}.
\] (5.3.3)

Following this we will consider estimation using conditional mean estimates based on adaptive model estimates. We assume in this section that \(B\) and \(D\) are known. We consider the following parameter estimates:
\[
\hat{A}_{k|k,\lambda} = \hat{J}_{k|k,\lambda} \hat{O}_{k|k,\lambda}^{-1} \quad \text{and} \quad \hat{C}_{k|k,\lambda} = \hat{T}_{k|k,\lambda} \hat{O}_{k|k,\lambda}^{-1}.
\] (5.3.4)

The model estimate is adaptively updated as follows:
\[
\hat{\lambda}_{k+1} = \lambda(\hat{A}_{k|k,\lambda}, B, \hat{C}_{k|k,\lambda}, D) \quad \text{and} \quad \hat{\lambda}_{k+1} = \hat{\lambda}_0, \ldots, \hat{\lambda}_{k+1}.
\] (5.3.5)

Suppose the persistently excitation condition associated with the signal generating system \(\lambda\) and its estimate \(\hat{\lambda}_k\) holds, so that:
\[
\lim_{k \to \infty} \sup_k \left( \frac{\hat{O}_{k|k,\lambda}}{k} \right)^{-1}
\] exists. (5.3.6)

5.3.1 Preliminary Convergence Result

**Theorem 5.3.1** Consider the linear system (5.2.1), (5.2.2) denoted by \(\lambda\). Suppose (5.2.16) and (5.2.17) hold. Also assume that conditional mean estimates based on the “correct model” are available. Assume the true model \(\lambda \in \Lambda\). Then
\[
\lim_{k \to \infty} \hat{A}_{k|k,\lambda}, \hat{C}_{k|k,\lambda} = A, C \quad \text{a.s.}
\] (5.3.7)
and the almost sure convergence rate is at least \((k^{-2} \rho(k))^{\frac{1}{2}}\).

**Proof:** We first proceed with the lemma result for \(\hat{A}_{k|k,\lambda}\). Simple manipulations of (5.2.6), (5.3.2) and (5.3.4) give the error term, as:
\[
\hat{A}_{k|k,\lambda} = \hat{A}_{k|k,\lambda} - A = E[k^{-1} B M_k | Y_k, \lambda] \\
\cdot \left( E[k^{-1} O_k | Y_k, \lambda] \right)^{-1}.
\] (5.3.8)

From Lemma 5.2.2 write that \(\phi_k := \rho(k)^{-\frac{1}{2}} M_k = \rho(k)^{-\frac{1}{2}} \sum_{i=1}^{k} \Delta M_i\), so that \(\lim_{k \to \infty} \phi_k = 0_{1 \times N} \) a.s. Now, \(\phi_k\) is bounded in \(L_2\) since
\[
E[\phi_k \phi_k^T] = E \left[ \rho(k)^{-\frac{1}{2}} w_{k+1} x_k w_{k+1}^T \rho(k)^{-\frac{1}{2}} \right]
\]
\[
E \left[ \rho(k)^{-1} E \left[ u^2_{k+1} | \mathcal{F}_k \right] E \left[ x_k^t x_k | \mathcal{F}_k \right] \right] \\
= \rho(k)^{-1} E \left[ x_k^t x_k \right] \\
< \infty.
\]

Here we have used the fact that for stable systems \( E[x_k^t x_k] < B_{\infty} \) for all \( k \) for some \( B_{\infty} < \infty \) so that \( E[x_k^t x_k | \mathcal{F}_k] < B_{\infty} \). Also \( E \left[ u^2_{k+1} | \mathcal{F}_k \right] = 1 \).

This is a uniform integrability condition, which together with the property that \( \lim_{k \to \infty} \phi_k = 0_{1 \times N} \) a.s. ensures convergence in conditional mean:

\[
E[\phi_k | \mathcal{Y}_k, \lambda] \to 0_{1 \times N} \text{ a.s.}
\]

(Also \( E[\phi_k] \to 0_{1 \times N} \) a.s.). Hence, \( E[k^{-1} M_k | \mathcal{Y}_k, \lambda] \to 0 \) a.s. at a rate \((k^{-2} \rho(k))^\delta\). This gives the convergence (and rate) result (5.3.7) for \( \hat{A}_{k|k, \lambda} \) under the excitation condition (5.3.6).

Similarly, the lemma convergence result holds for \( \hat{C}_{k|k, \lambda} \).

\( \square \)

**Remarks**

1. Optimal finite-dimensional filters for \( \hat{O}_{k|k, \lambda}, \hat{J}_{k|k, \lambda} \) and \( \hat{T}_{k|k, \lambda} \) are known and require \( O(N^3) \) calculations per time instant [22]. Suboptimal filters can be implemented for \( \hat{O}_{k|k, \lambda}, \hat{J}_{k|k, \lambda} \) and \( \hat{T}_{k|k, \lambda} \) from Kalman filter state estimates. For example, \( \hat{O}_{k|k, \lambda}^{mb} = \sum_{i=1}^{k} \tilde{x}_{i-1} \tilde{x}_{i-1} \). Convergence results when suboptimal estimates are used is neither excluded or included by this theory.

We proceed to consider in the next section the case when conditional mean estimates based on an adaptive model estimate are used.

### 5.3.2 Global Convergence Result

**Notation:** To make use of the convergence rate results of [29] we introduce an \( M \) bounded notation. We say \( \epsilon \) is \( O_M(k^{-1/2}) \) if \( \epsilon \) is \( M \) bounded in the sense that \( \sup_{k \geq 1} E \left[ \frac{\epsilon}{k^{-1/2}} \right]^{1/2} = M_k(q) < \infty \) for all \( 1 \leq q < \infty \).

**Theorem 5.3.2** Consider the linear system (5.2.1),(5.2.2) denoted by \( \lambda \). Suppose (5.2.16) and (5.2.17) hold. Consider a sequence of estimated models \( \hat{A}_k \) adaptively updated by previous parameter estimates so that \( \hat{A}_{k+1} = \lambda(\hat{A}_{k|k, \lambda}, \hat{B}_k, \hat{C}_{k|k, \lambda}, D) \). We suppose that \( \hat{A}_k \) is
persistently exciting, along with λ in that (5.3.6) holds. Then

$$\lim_{k \to \infty} \tilde{A}_{k|k, \lambda_k}, \tilde{C}_{k|k, \lambda_k} = A, C \quad a.s.$$  \hspace{1cm} (5.3.9)

and the errors \( \tilde{A}_{k|k, \lambda_k} - A \) and \( \tilde{C}_{k|k, \lambda_k} - C \) are \( O_M(k^{-1/2}) \).

**Proof:** The following proof requires that the state process \( \{x_k\} \) be asymptotically ergodic. The persistence of excitation conditions (5.2.17) and (5.3.6) imply this. We also require that the filters for \( \tilde{J}_{k|k, \lambda_k}, \tilde{O}_{k|k, \lambda_k} \) and \( \tilde{T}_{k|k, \lambda_k} \) are exponentially forgetting.

First we consider estimation of \( A \) only. Simple manipulation of (5.3.4) gives the recursions for the estimate \( \tilde{A}_{k|k, \lambda_k} \) as follows:

$$\tilde{A}_k = \tilde{A}_{k-1} + \frac{1}{k} (\Delta J_k - \tilde{A}_{k-1} \Delta O_k) \left( \frac{1}{k} \tilde{O}_{k|k, \lambda_k} \right)^{-1},$$  \hspace{1cm} (5.3.10)

where we denote \( \tilde{A}_{k|k, \lambda_k} \) by the shorthand notation \( \tilde{A}_k \) and define \( \Delta J_k := \tilde{J}_{k|k, \lambda_k} - \tilde{J}_{k-1|k-1, \lambda_{k-1}} \) and \( \Delta O_k := \tilde{O}_{k|k, \lambda_k} - \tilde{O}_{k-1|k-1, \lambda_{k-1}} \). Note that if \( \tilde{A}_k \) leaves \( M_{st} \) then projection back onto \( M_{st} \) is performed, where \( M_{st} \) is the set of stable matrices.

Convergence of recursion (5.3.10) can be shown by considering an associated ordinary differential equation (ODE). We follow the techniques introduced and discussed in [51, 50, 42, 29]. That is,

$$\frac{d\tilde{A}^c(\tau, k)}{d\tau} = R^{-1}(\tau, k)f(\tilde{A}(\tau, k), k)$$

$$\frac{dR(\tau, k)}{d\tau} = G(\tilde{A}(\tau, k), k), \quad R(0, k) \geq \delta I,$$  \hspace{1cm} (5.3.11)

where \( k \) here is a fixed parameter and \( \tilde{A}^c(\tau, k) := \text{col vec}(\tilde{A}(\tau, k)) \), where for an arbitrary matrix \( A \) we define \( \text{col vec}(A) := [a_{11}, \ldots, a_{1N}, a_{21}, \ldots, a_{2N}, \ldots, a_{NN}]' \). We also define a model set \( D_M \) which is a compact set for which (5.3.10) is stable. \( D_M \) will be a domain of attraction for (5.3.11). If \( \tilde{A}(\tau, k) \) leaves \( D_M \) then we project back onto the stability set \( D_M \).

Also, with \( \tilde{A}(\tau, k) \) abbreviated as \( \tilde{A} \) we define \( f(\tilde{A}, k) \) and \( G(\tilde{A}, k) \) as follows:

$$f(\tilde{A}, k) := \text{col vec}(E[\Delta \tilde{J}_{k|\tilde{A}} - \Delta \tilde{O}_{k|\tilde{A}}])$$

$$G(\tilde{A}, k) := I_N \otimes E[\Delta \tilde{O}_{k|\tilde{A}}],$$  \hspace{1cm} (5.3.12)

where \( I_N \) is the identity matrix of size \( N \times N \) and \( \otimes \) is the Kronecker product. Notice that \( G \) is block diagonal.
Consider the function,
\[
W(\bar{A}, k) = E \left[ E \left[ \sum_{\ell=1}^{k} ||x_{\ell} - \bar{A}x_{\ell-1}||^2 \bigg| \bar{A}, \mathcal{Y}_k \right] \right] - E \left[ \sum_{\ell=1}^{k-1} ||x_{\ell} - \bar{A}x_{\ell-1}||^2 \bigg| \bar{A}, \mathcal{Y}_{k-1} \right] \right]. \tag{5.3.13}
\]

It follows from classical expectation results, including that \(E[\mathbb{E}[X|A_2]|A_1] = E[X|A_1]\) when \(A_1 \subset A_2\), that \(W(\bar{A}, k) = E[||x_k - \bar{A}x_{k-1}||^2] \geq 0\). Under asymptotic ergodicity and certain smoothness conditions the differentiation w.r.t. \(\bar{A}\) and the expectation operations can be interchanged, for \(k\) sufficiently large. Hence, for large \(k\),
\[
\frac{dW(\bar{A}(\tau, k), k)}{d\bar{A}^c(\tau, k)} = -f(\bar{A}(\tau, k), k) \tag{5.3.14}
\]
and it then follows that for large \(k\)
\[
\frac{dW(\bar{A}(\tau, k), k)}{d\tau} = \frac{dW(\bar{A}(\tau, k), k)}{d\bar{A}^c(\tau, k)} \frac{d\bar{A}^c(\tau, k)}{d\tau} = -f(\bar{A}(\tau, k), k) R(\tau, k)^{-1} f(\bar{A}(\tau, k), k). \tag{5.3.15}
\]

It follows from [73, Page 63] that \(O_k\) is a solution of the Lyapunov equation and hence is nonnegative definite. The persistence of excitation condition (5.2.17) further implies that the nonnegative definite \(O_k\) must be positive definite. Hence, \(\frac{dW(\bar{A}, k)}{d\tau} < 0\) for sufficiently large \(k\) because \(R(\tau, k)\) is positive definite for all \(\tau\) and \(k\). It follows that \(W(\bar{A}, k)\) is a Lyapunov function in \(D_M\) for sufficiently large \(k\).

It follows from Lyapunov’s direct method and equation (5.3.14) that \(\bar{A}(\tau, k)\) converges to the set \(\{\bar{A} | \lim_{k \to \infty} f(\bar{A}, k) = 0\} \subset D_M\) or to the boundary of \(D_M\) a.s.

Under asymptotic ergodicity we have in the limit \(k \to \infty\) the set \(\{\bar{A} | \lim_{k \to \infty} f(\bar{A}, k) = 0\}\) is equivalent to the set \(\{\bar{A} | \lim_{k \to \infty} E \left[ \tilde{J}_k^{i|k, \bar{A}_k|} - \tilde{A}_k^{i|k, \bar{A}_k|} \right] = 0\}\). It follows from Theorem 5.3.1 that the true model \(A\) is a member of this set.

To show that convergence is global we convert the state space representation into the observable canonical form because the \((A, B, C, D)\) state-space representation of linear system is a non-unique representation. Convergence occurs to solutions of the \(N^2\) simultaneous equations \(\frac{dW(\bar{A}, k)}{d\alpha_{ij}} = E[\tilde{x}_i^{j|k}|x_{\ell-1}^{j|k} = \sum_{n=1}^{N} \alpha_{ij} x_{\ell-1}^n] = 0\) for \(1 \leq i, j \leq N\) (Actually only \(N\) distinct equations). The observable canonical form has only \(N\) free variables and will have only one solution to these simultaneous equations (if there is one). The persistence of excitation conditions imply that there is a solution, hence \(\bar{A}\) converges uniquely to the true value of \(A\) in the observable canonical form.
Under the asymptotic ergodicity conditions, the ODE convergence result translates across to imply convergence of difference equation (5.3.10), hence in the limit as \( k \to \infty \), \( \hat{A}_{k|k|, k} \) (or \( \hat{A}_k \) in (5.3.10)) converges to the true \( A \).

To establish convergence rate results we appeal to Theorem 4.1 of [29]. In the notation of [29], we have \( x = \text{col vec}(\bar{A}), \phi = \left( \text{col vec}(\Delta \bar{J}_{k|A}), \text{col vec}(\Delta \bar{O}_{k|A}) \right) \),

\[
Q(\phi) = \text{col vec} \left( (\Delta \bar{J}_{k|A} - \bar{A} \Delta \bar{O}_{k|A}) (\Delta \bar{O}_{k|A})^{-1} \right),
\]

\( D_0 = \{ \Delta \bar{J}_{k|A}, \Delta \bar{O}_{k|A} : \bar{A} \in D_M \} \), and

\( D_0 \cap D_M \) is a compact domain invariant for (5.3.11). Note from the above convergence result for (5.3.11) it is implied that some \( D_{inv} \) does indeed exist.

This notation together with the persistence of excitation conditions satisfy the conditions of Theorem 4.1 of [29] (perhaps with slight modification of the recursions) and hence it follows directly from application of Theorem 4.1 of [29] that \( \hat{A}_{k|k|, k} = O_M(k^{-1/2}) \).

We proceed to prove convergence when simultaneously estimating \( A \) and \( C \).

Simple manipulation of (2.11) gives

\[
\bar{C}_k = \bar{C}_{k-1} + \frac{1}{k} (\Delta T_k - \bar{C}_{k-1} \Delta \bar{O}_k) (\frac{1}{k} \bar{O}_{k|k|, k} \bar{A} \bar{A}^{-1}, \bar{A} \bar{A}^{-1} \Delta \bar{O}_k)^{-1}, \tag{5.3.16}
\]

where \( \bar{C}_k \) is shorthand notation for \( \bar{C}_{k|k|, k} \) and \( \Delta T_k := \bar{T}_{k|k|, k} - \bar{T}_{k-1|k-1|, k} \). Now consider the ODE associate with the \( A \) and \( C \) recursions

\[
\frac{d\bar{\theta}(\tau, k)}{d\tau} = R^{-1}(\tau, k) f(\bar{\theta}(\tau, k), k)
\]

\[
\frac{dR(\tau, k)}{d\tau} = G(\bar{\theta}(\tau, k), k), \quad R(0, k) \geq \delta I, \tag{5.3.17}
\]

where \( k \) is here a fixed parameter, with \( \bar{A}(\tau, k) \) abbreviated as \( \bar{A} \) we define \( f(\bar{\theta}, k) \) and \( G(\bar{\theta}, k) \) as follows

\[
f(\bar{\theta}, k) := \left[ \begin{array}{c}
\text{col vec}(E[\Delta \bar{J}_{k|\bar{\theta}} - \bar{A} \Delta \bar{O}_{k|\bar{\theta}}]) \\
\text{col vec}(E[\Delta \bar{T}_{k|\bar{\theta}} - \bar{C} \Delta \bar{O}_{k|\bar{\theta}}])
\end{array} \right]
\]

\[
G(\bar{\theta}, k) := I_{N+1} \otimes E[\Delta \bar{O}_{k|\bar{\theta}}] \tag{5.3.18}
\]

where \( \bar{\theta}(\tau, k) = \left[ \begin{array}{c}
\text{col vec}(\bar{A}(\tau, k)) \\
\text{col vec}(\bar{C}(\tau, k))
\end{array} \right] \) with \( \bar{A}(\tau, k) \) abbreviated as \( \bar{A} \). Consider the Lyapunov function,

\[
\bar{W}(\bar{\theta}, k) = E \left[ \sum_{\ell=1}^{k} ||x_{\ell} - \bar{A}x_{\ell-1}||^2 \left| \bar{A}_Y \right| + E \left[ \sum_{\ell=1}^{k-1} ||x_{\ell} - \bar{A}x_{\ell-1}||^2 \left| \bar{A}_Y \right| \right]
\]

\[
+ E \left[ \sum_{\ell=1}^{k} ||y_{\ell} - \bar{C}x_{\ell}||^2 \left| \bar{A}_Y \right| + E \left[ \sum_{\ell=1}^{k-1} ||y_{\ell} - \bar{C}x_{\ell}||^2 \left| \bar{A}_Y \right| \right] \right]. \tag{5.3.19}
\]
In a similar manner to the above case for estimation of \( A \), it follows that \( \bar{W}(\bar{\theta}(\tau,k),k) \geq 0 \) for sufficiently large \( k \) and that \( \frac{d\bar{W}(\bar{\theta}(\tau,k),k)}{d\bar{\theta}(\tau,k)} = -f(\bar{\theta}(\tau,k),k) \), and thus \( \frac{d\bar{W}(\bar{\theta}(\tau,k),k)}{dt} < 0 \) for sufficiently large \( k \). Hence, \( \bar{W}(\bar{\theta}(k)) \) is a Lyapunov function for sufficiently large \( k \).

It follows from Lyapunov’s direct method that the \( A \) and \( C \) estimates converge to the set

\[
\left\{ \bar{\mu}, \bar{C} \mid \lim_{k \to \infty} \frac{d\bar{W}(\bar{\theta}(\tau,k),k)}{dt} = 0 \right\}.
\]

The set \( \left\{ \bar{\mu}, \bar{C} \mid \lim_{k \to \infty} \frac{d\bar{W}(\bar{\theta}(\tau,k),k)}{dt} = 0 \right\} \) is equivalent to the sets \( \left\{ \bar{A} \mid \lim_{k \to \infty} E\left[ \hat{J}_{k|k} - \bar{A}\hat{\theta}_{k|k} \right] = 0 \right\} \) and \( \left\{ \bar{C} \mid \lim_{k \to \infty} E\left[ \hat{T}_{k|k} - \bar{A}\hat{\theta}_{k|k} \right] = 0 \right\} \).

It follows from Theorem 5.3.1 that the true model parameters \( A \) and \( C \) are members of these convergence sets. In a similar manner to the earlier results, global convergence and convergence rates follow.

\[ \square \]

Remarks

1. When implementing the filters for \( J_k \) and \( O_k \) the noise variances \( B \) and \( D \) may need to be increased to include the induced noise effect of parameter estimate errors.

2. Here we have considered only adaptive estimation of \( A \) and \( C \). Global convergence of adaptive estimation of \( B \) and \( D \) is neither excluded or included by this theory.

5.4 Hybrid Systems

In this section we introduce a hybrid linear system model with states \( x_k \) in a continuous range, and other states \( X_k \) in a discrete set. Our time parameter set is the non-negative integers \( \mathbb{Z}^+ = \{0,1,2,\ldots\} \).

On a probability space \((\Omega, \mathcal{F}, P)\) we suppose that we have a discrete-time, discrete-valued homogeneous Markov chain \( \{X_\ell\}, \ell \in \mathbb{Z}^+ \) taking on values in a finite set and that we have a discrete-time continuous-valued stochastic process \( \{x_\ell\}, \ell \in \mathbb{Z}^+ \) taking on values in \( \mathbb{R}^N \). Let \( \mathcal{G}_k = \sigma\{x_0, x_1, \ldots, x_k, X_0, \ldots, X_k\} \) denote the filtration generated by \( x_k \) and \( X_k \).

First consider the Markov chain, which is assumed independent of the continuous-valued state \( x_k \). As pointed out in [18], without loss of generality we can take the state space of the Markov chain to be the set \( S = \{e_1, e_2, \ldots, e_N\} \) of unit vectors in \( \mathbb{R}^N \). Here \( e_i = (0,0,\ldots,1,\ldots,0)' \in \mathbb{R}^N \). That is, at each \( t, X_\ell \in S \). The transition matrix is

\[
A = (A_{ij}), \quad 1 \leq i, j \leq N.
\]
\[ A_{ji} = P(X_t = e_j | X_{t-1} = e_i) \]

so that

\[ E[X_{k+1} | \mathcal{G}_{k-1}] = AX_k \]  \hspace{1cm} (5.4.2)

The dynamics of \( X_k \) are given by the state equation, see [18],

\[ X_{k+1} = AX_k + V_{k+1}, \quad X_0 \in S \]

where \( V_{k+1} := X_{k+1} - AX_k \) is a martingale increment on \( \mathcal{G}_k \), as in [18], since,

\[ E[V_{k+1} | \mathcal{G}_k] = 0 \in R^N. \]

The continuous-valued state is generated by a hybrid (Markov jump) linear system with vector state \( x_k \), and driven by white noise and the discrete state \( X_k \). This continuous-valued state, in turn, is observed in white noise.

For \( k \in \mathbb{Z}^+ \), \( A(e_i) \in \mathbb{R}^{N \times N} \) and \( B(e_i) \in \mathbb{R}^{N \times 1} \) for \( i = 1, \ldots, N \) and we suppose that

\[ x_{k+1} = A(X_k)x_k + B(X_k)w_{k+1}, \quad x_0 \in \mathbb{R}^{N \times 1}. \]  \hspace{1cm} (5.4.3)

where \( \{w_l\}, l \in \mathbb{Z}^+ \) is an independent \( N(0, 1) \) scalar random variable.

The state process \( x_k \) is observed indirectly via the scalar observation process \( \{y_l\}, l \in \mathbb{Z}^+ \). Assume \( y_k \) takes on values in \( R \). For \( k \in \mathbb{Z}^+ \), \( C(e_i) \in \mathbb{R}^{1 \times N} \) and \( D(e_i) \in \mathbb{R} \) for \( i = 1, \ldots, N \) and we suppose that

\[ y_k = C(X_k)x_k + D(X_k)v_k, \quad y_k \in R. \]
\[ z_k = C(X_k) + D(X_k)v_k, \quad z_k \in R \]  \hspace{1cm} (5.4.4)

where \( \{v_l\}, l \in \mathbb{Z}^+ \) is an independent \( N(0, 1) \) scalar random variable and \( \{v_l\}, l \in \mathbb{Z}^+ \) is an independent \( N(0, 1) \) scalar random variable. Assume that \( \{v_l\}, \{w_l\}, x_0 \) and \( X_0 \) are independent.

Denote the model (5.4.3), (5.4.3) and (5.4.4) by

\[ \lambda^H = \lambda^H(A, B, C, D, A, C, D, x_0, X_0). \]  \hspace{1cm} (5.4.5)

Let \( \mathcal{F}_k \) denote the complete filtration generated by \( \sigma\{x_0, x_1, \ldots, x_k, X_0, \ldots, X_k y_0, \ldots, y_k, z_0, \ldots, z_k\} \). Also, let \( \mathcal{Y}_k \) by the complete filtration generated by \( \sigma\{y_0, \ldots, y_k, z_0, \ldots, z_k\} \) Also, recall the property of the indicator vectors \( X_k \), that
non-linear functions can be written as linear functions, that is,

\[ A(X_k) = \sum_{i=1}^{N} A(e_i)X_k^i \]

where \( X_k^i \) is the \( i \)th element of \( X_k \).

### 5.4.1 Measure Change

All the processes are defined on a probability space \((\Sigma, \mathcal{F}, P)\); Here we define a new probability measure \( \bar{P} \) such that under \( \bar{P} \):

1) \( \{x_k\} \) is a sequence of i.i.d real \( N(0,1) \) with density \( \psi \).
2) \( \{y_k\} \) is a sequence of i.i.d real \( N(0,1) \) with density \( \phi \).
3) \( \{z_k\} \) is a sequence of i.i.d real \( N(0,1) \) with density \( \gamma \).

The new probability measure \( \bar{P} \) can be defined by setting the restriction of the Radon-Nikodym derivative \( d\bar{P}/dP \) to \( \mathcal{F}_k \) equal to \( \Pi_k \) [18] where

\[
\Pi_k = \prod_{i=1}^{k} \pi_i \text{ for } k \geq 1 \quad (5.4.6)
\]

\[
\pi_{k+1} = \frac{\gamma(\frac{\hat{C}(X_k+1,x_i)}{D(X_k)}) \phi(\frac{\hat{C}(X_k+1,x_i)}{D(X_k)}) \psi(\frac{\hat{C}(X_k+1,x_i)}{D(X_k)})}{B(X_k) \phi(x_{k+1})} \quad (5.4.7)
\]

Let \( g : R \to R \) be a “test function”, (i.e. measurable with compact support). Recall from [18] a version of Bayes’ theorem states that for the \( \mathcal{F} \) adapted sequence \( \{(X_k, e_i)g(x_k)\} \)

\[
E[(X_k, e_i)g(x_k)|Y_k] = \frac{\bar{E}[\Pi_k(X_k, e_i)g(x_k)|Y_k]}{\bar{E}[\Pi_k|Y_k]} \quad (5.4.8)
\]

where \( \bar{E}[\cdot] \) denotes expectation in \( \bar{P} \) and \( \overline{\Pi}_k \) denotes the reverse measure change, see [18].

### 5.4.2 Preliminary Results for Parameter Estimation

Estimation of \( A, B, C \) and \( D \) is considered here. Estimation of \( A, C \) and \( D \) is examined in previous chapters and those results carry over to this system. From (5.4.3), by post-multiplication by \( X_{\ell-1}^i x_{\ell-1}^j \) for \( i = 1, \ldots, N \) and summation over \( \ell = 1, \ldots, k \), we have that

\[
\sum_{\ell=1}^{k} x_{\ell-1}^i x_{\ell-1}^j X_{\ell-1}^i = \sum_{j=1}^{N} \left( A(e_j) \sum_{\ell=1}^{k} x_{\ell-1}^i x_{\ell-1}^j X_{\ell-1}^i + B(e_j) \sum_{\ell=1}^{k} w_{\ell} x_{\ell-1}^i X_{\ell-1}^i \right) \quad \text{for } i = 1, \ldots, N. \quad (5.4.9)
\]
Now noting that \( X_i^j X_j^i = 0 \) for all \( i \neq j \), we obtain \( N \) equations:

\[
\sum_{\ell=1}^{k} x_{\ell} x'_{\ell-1} X_i^\ell = A(e_i) \sum_{\ell=1}^{k} x_{\ell-1} x'_{\ell-1} X_i^\ell + B(e_i) \sum_{\ell=1}^{k} w_{\ell} x'_{\ell-1} X_i^\ell \quad \text{for} \quad i = 1, \ldots, N.
\]  

(5.4.10)

Now write

\[
\mathcal{J}_{k,i} = \sum_{\ell=1}^{k} x_{\ell} x'_{\ell-1} X_i^\ell, \quad \mathcal{O}_{k,i} = \sum_{\ell=1}^{k} x_{\ell-1} x'_{\ell-1} X_i^\ell, \quad \text{and} \quad \mathcal{M}_{k,i} = \sum_{\ell=1}^{k} w_{\ell} x'_{\ell-1} X_i^\ell.
\]

(5.4.11)

Substitution gives

\[
\mathcal{J}_{k,i} = A(e_i) \mathcal{O}_{k,i} + B(e_i) \mathcal{M}_{k,i} \quad \text{for} \quad i = 1, \ldots, N.
\]  

(5.4.12)

Thus a reasonable estimate for \( A(e_i) \) is

\[
\hat{A}_k(e_i) = \mathcal{J}_{k,i} (\mathcal{O}_{k,i})^{-1} \quad \text{for} \quad i = 1, \ldots, N.
\]  

(5.4.13)

The error in these estimates is

\[
\hat{A}_k(e_i) - A(e_i) = B(e_i) \mathcal{M}_{k,i} (\mathcal{O}_{k,i})^{-1} \quad \text{for} \quad i = 1, \ldots, N.
\]  

(5.4.14)

Similarly, from (5.4.4) by post-multiplication by \( X_i^\ell x'_{\ell-1} \) for \( i = 1, \ldots, N \) and summation over \( \ell = 1, \ldots, k \), we have that

\[
\sum_{\ell=1}^{k} y_{\ell} x'_{\ell-1} X_i^\ell = C(e_i) \sum_{\ell=1}^{k} x_{\ell-1} x'_{\ell-1} X_i^\ell + D(e_i) \sum_{\ell=1}^{k} v_{\ell} x'_{\ell-1} X_i^\ell \quad \text{for} \quad i = 1, \ldots, N.
\]  

(5.4.15)

Now write

\[
\mathcal{T}_{k,i} = \sum_{\ell=1}^{k} y_{\ell} x'_{\ell-1} X_i^\ell, \quad \text{and} \quad \mathcal{V}_{k,i} = \sum_{\ell=1}^{k} v_{\ell} x'_{\ell-1} X_i^\ell.
\]

(5.4.16)

Then (5.4.15) can be written \( \mathcal{T}_{k,i} = C(e_i) \mathcal{O}_{k,i} + D(e_i) \mathcal{V}_{k,i} \). Therefore, a reasonable estimate for \( C(e_i) \) is

\[
\hat{C}_k(e_i) = \mathcal{T}_{k,i} (\mathcal{O}_{k,i})^{-1}
\]

(5.4.17)

and the error in this estimation is \( \hat{C}_k(e_i) - C(e_i) = D(e_i) \mathcal{V}_{k,i} (\mathcal{O}_{k,i})^{-1} \).

### 5.4.3 Almost Sure Convergence

In this subsection we discuss the convergence of these estimates. Before proceeding to our convergence results we first state a stability condition.
The hybrid linear system is stable if the following condition holds

$$\max_{i=1, \ldots, N} |\sigma_{\max}(A(e_i))| < 1$$  \hspace{1cm} (5.4.18)

where $\sigma_{\max}(A)$ is the largest eigenvalue of the matrix $A$. Note that this is a sufficient but not necessary condition for stability.

We also note that the hybrid system (5.4.3),(5.4.4) is persistently excited (except for the trivial case when $B(e_i) = 0_{N \times 1}$ for $i = 1, \ldots, N$, where $0_{N \times 1}$ is the $(N \times 1)$ vector of all zeros) in that the following condition is satisfied

$$\lim_{k \to \infty} \sup_{k} \left( \frac{1}{k} O_k^m \right)^{-1} < \infty \quad \text{for} \quad i = 1, \ldots, N$$  \hspace{1cm} (5.4.19)

where $O_k^m$ is the $mn$th element of $O_k$.

Let us introduce a positive, non-decreasing scalar function, $\rho(k)$, such that

$$\lim_{k \to \infty} \sum_{i=0}^{k} \rho(i)^{-1} = \lambda(w) < \infty.$$ An example of such a $\rho(k)$ is $\rho(k) = \max(1, k \ln k (\ln \ln k)^{\alpha})$, for any $\alpha > 1$.

The following lemma now holds.

**Lemma 5.4.1** Consider the hybrid linear system (5.4.3)(5.4.4) with states $X_k$ and $x_k$, $\mathcal{Y}_k$-measurable. If the system is stable in that (5.4.18) holds, and all states are persistently excited in that (5.4.19) holds, then

$$\lim_{k \to \infty} \tilde{A}_k(e_i), \tilde{C}_k(e_i) = A(e_i), C(e_i) \quad \text{a.s. for} \quad i = 1, \ldots, N$$

and the almost sure convergence rate is guaranteed to be $(k^{-2} \rho(k))^{1/2}$.

**Proof**: Define $\phi_{k,i} := \sum_{j=1}^{k} \rho(j)^{-1/2} \Delta M_{k,i}^j$ where $\Delta M_{k,i}^j = M_{k,i} - M_{k-1,i}$. Hence $\phi_{k,i} = \sum_{j: X_{j-1} = e_i} w_j x_{j-1}$. Now $\phi_{k,i}$ are vectors whose elements are martingales on $(P, \mathcal{G}, \Sigma)$ since

$$E[\phi_{k,i} | \mathcal{F}_{k+1}] = E[\phi_{k-1,i} + w_k x_{k-1} X_{k-1}^i | \mathcal{F}_{k+1}] = \phi_{k-1,i}.$$ Also the elements of $\phi_{k,i}$ are bounded in $L_2$ since

$$E[(\phi_{k,i}^j)^2] = E \left[ \sum_{\ell=1}^{k} \rho(\ell)^{-1} \Delta M_{\ell,i}^j \Delta M_{\ell,i}^j \right]$$

$$= E \left[ \sum_{\ell: X_{\ell-1} = e_i} \rho(\ell)^{-1} E \left[ (x_{\ell-1}^j)^2 w_{\ell}^2 | \mathcal{F}_{\ell-1} \right] \right]$$

$$= E \left[ \sum_{\ell: X_{\ell-1} = e_i} \rho(\ell)^{-1} E \left[ |x_{\ell-1}^j|^2 \right] \right]$$

$$= \eta(w)^j < \infty.$$
Here \( \phi_{k,i}^j, \Delta M_{\ell,i}^j \) and \( x_{\ell-1}^j \) are the \( j \)th elements of \( \phi_{k,i}, \Delta M_{\ell,i} \) and \( x_{\ell-1} \), respectively, and we have used the fact that for stable hybrid systems \( E[|x_{\ell-1}^j|^2] < B_\infty \) for all \( \ell,j \) for some \( B_\infty < \infty \) and that \( E[w_i^j|F_{\ell-1}] = 1 \).

Since the elements of \( \phi_{k,i} \) are martingales and bounded in \( L_2 \) then standard martingale convergence results [59, 64] give that the elements of \( \phi_{k,i} \) converge to finite random variables for \( i = 1, \ldots, N \). Hence, it follow by the Kronecker lemma [53, 64] that

\[
\lim_{k \to \infty} \rho(k)^{-1/2} M_{k,i} = 0_{1 \times N} \quad a.s. \quad \text{for } i = 1, \ldots, N.
\]

Under the excitation condition, and using the result that the limit of products is the product of limits (when these limits exist), then the error

\[
\tilde{\mathcal{A}}(e_i) := \hat{\mathcal{A}}(e_i) - A(e_i) = B \left( \frac{1}{k} M_{k,i} \right) \left( \frac{1}{k} O_{k,i} \right)^{-1} \to 0_{N \times N}
\]

and the rate of convergence is \( (k^{-2/3}\rho(k))^2 \). The result for \( C \) follows similarly.

\[ \square \]

Remarks

1. Partial convergence results can be shown when the persistence of excitation condition holds for some, but not all \( i \).

2. Estimators for \( A, C \) and \( D \) are given in Chapter 4.

3. The stability condition (5.4.18) is a sufficient condition for convergence but may be too restrictive.

5.4.4 Conditional Mean Estimates

Consider again the system (5.4.3), (5.4.4). Let us denote associated conditional mean estimates based on the signal generating system \( \lambda^H \), also termed the “correct” model, and associated conditional mean estimates based on a possibly time varying estimate \( \hat{\lambda}^H_1 = \hat{\lambda}^H, \ldots, \hat{\lambda}^H_k \), also termed the “incorrect” model, as, respectively,

\[
\mathcal{J}_{k,i|k,\hat{\lambda}^n} = E[\mathcal{J}_{k,i}|\mathcal{Y}_k, \hat{\lambda}^H] \quad \mathcal{O}_{k,i|k,\hat{\lambda}^n} = E[O_{k,i}|\mathcal{Y}_k, \hat{\lambda}^H] \quad (5.4.20)
\]
and
\[
\tilde{\mathcal{K}}_{k,i}[k,\lambda^u_t] = \mathcal{E}[\mathcal{J}_{k,i}[\mathcal{Y}_k,\lambda^H_t] \quad \tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t] = \mathcal{E}[\mathcal{O}_{k,i}[\mathcal{Y}_k,\lambda^H_t]]
\] (5.4.21)

In this section we consider the parameter estimates, with the true model known or with model estimates available, respectively,
\[
\tilde{A}_{k|k,\lambda^u_t}(e_i) = \mathcal{J}_{k,i}[k,\lambda^u_t] (\tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t])^{-1}, \quad \tilde{\mathcal{C}}_{k|k,\lambda^u_t}(e_i) = \mathcal{J}_{k,i}[k,\lambda^u_t] (\tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t])^{-1}
\] (5.4.22)

and
\[
\tilde{\mathcal{A}}_{k|k,\lambda^u_t}(e_i) = \mathcal{J}_{k,i}[k,\lambda^u_t] (\tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t])^{-1}, \quad \tilde{\mathcal{C}}_{k|k,\lambda^u_t}(e_i) = \mathcal{J}_{k,i}[k,\lambda^u_t] (\tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t])^{-1}
\] (5.4.23)

We introduce a second persistence of excitation condition associated with the model \(\lambda^H_t\) and its estimate \(\tilde{\lambda}^H_t\) as
\[
\lim_{k \to \infty} \sup_{i,j} \sup \left( \frac{1}{k} \tilde{\mathcal{O}}_{k,i}[k,\lambda^u_t] \right)^{-1} < \infty
\] (5.4.24)

Then the following theorem holds.

**Theorem 5.4.1** Consider the hybrid linear process (5.4.3), (5.4.4) denoted by \(\lambda^H_t\). Suppose that (5.4.19) and (5.4.18) hold. Also assume that the conditional mean estimates based on the “correct model” are available. Then
\[
\lim_{k \to \infty} \tilde{A}_{k|k,\lambda^u_t}(e_i), \tilde{\mathcal{C}}_{k|k,\lambda^u_t}(e_i) = A(e_i), C(e_i) \quad \text{a.s. for } i = 1, \ldots, N.
\] (5.4.25)

The almost surely convergence rate is guaranteed is \((k^{-2}\rho(k))^{\frac{1}{2}}\).

**Proof** : Consider the error term (5.4.14) when conditional mean estimates are used.
\[
\tilde{A}_k(e_i) := A(e_i) - \tilde{A}_k(e_i) = \mathcal{E}[k^{-1}\mathcal{M}_k,i][\mathcal{Y}_k,\lambda^H_t] - \mathcal{E}[k^{-1}\mathcal{O}_{k-1,i}[\mathcal{Y}_k,\lambda^H_t]]_{\text{diag}}
\]

Consequently, in addition to \(\phi_{k,i} = \rho(k)^{-\frac{1}{2}} M_{k,i}\) satisfying \(\lim_{k \to \infty} \phi_{k,i} = 0_{1 \times N} \) a.s. Now parrelling the derivation in Lemma 5.4.1 we have the following result
\[
\mathcal{E}[(\phi_{k,i}^j)^2] = (k^{-1}\ln(k)ln(k))^a \sum_{i \in \epsilon} \mathcal{E} \left( E^2 \left[ |x_{t-1}|^2 w^2_{t-1 \epsilon} \right] \right) < \infty.
\]
This is a uniform integrability condition, which together with the property that
\[
\lim_{k \to \infty} \phi_{k,i} = 0_{1 \times N} \text{ a.s.} \quad \text{ensures convergence in conditional mean:}
\]
\[
E[\phi_{k,i}|Y_k, \lambda^H] \to 0_{1 \times N} \text{ a.s.}
\]
The convergence (and rate) result (5.4.25) for \( \hat{C}(e_i) \) follows under the excitation condition (5.4.24).

\[\square\]

**Remarks**

1. Presently, calculation of the conditional mean estimates \( \hat{J}_{k,i|k}, \hat{\lambda}^H_h \), \( \hat{O}_{k,i|k}, \hat{\lambda}^H_h \) and \( \hat{T}_{k,i|k}, \hat{\lambda}^H_h \) is prohibitively computationally difficult. Filters for state estimation require computational effort which is exponential in \( k \), see [19]. Filters for the conditional mean estimates require more effort and also must be implemented for reasonably large \( k \) when used for parameter estimation.

2. Suboptimal estimation of \( \hat{J}_{k,i|k}, \hat{\lambda}^H_h \), \( \hat{O}_{k,i|k}, \hat{\lambda}^H_h \) and \( \hat{T}_{k,i|k}, \hat{\lambda}^H_h \), including batch processing, may result in practical parameter estimation techniques but is not considered here.

3. The \( A(e_i), C(e_i) \) representation of linear systems is non-unique. Estimation can only occur to an equivalence class. In terms of the Theorem 5.4.1 result, the true system is one system in the equivalence class.

**5.5 Simulations**

In this section, simulation studies are presented for a selection of the estimation schemes.

**Ex 1:** (Adaptive estimation) A 20000 point, 2-state linear system (5.2.1), (5.2.2) is generated with parametric values \( A = \begin{bmatrix} 0 & -0.8 \\ 1 & -0.1 \end{bmatrix} \), \( B = [1, 0]' \), \( C = [1, 0.2]' \) and \( D = 0.01 \). This system is in companion form and has eigenvalues of \(-0.0500 \pm 0.8930i\). In this simulation, \( C \) and \( D \) are assumed known and \( A \) and \( B \) are estimated by (5.3.2) and (5.3.4). Our random initial guess for the model is, \( \hat{A}_0 = \begin{bmatrix} 0.9347 & 0.5194 \\ 0.3835 & 0.8310 \end{bmatrix} \) which has eigenvalues of 1.3321 and 0.4335. The model estimate \( \hat{\lambda}_k \) used in the \( J_k \), \( O_k \) filters is updated at each time instant by the new parameter estimates \( \hat{A}_{k|k, \hat{\lambda}_k} \). Note that the estimated system is transformed into companion form at each time step. Figure 5.1 shows convergence of \( \hat{a}_{21} \). The parameter
estimates for the other elements of $A$ and other system parameters converged similarly. The final estimated system had eigenvalues of $-0.0361 \pm 0.8810i$, which compares well with the generating system. Note that the parameter estimators are not turned on until after 1000 points to allow the $J_k$ and $O_k$ filters to forget initial conditions.

**Ex 2:** (More Noise) The same process is generated with $D = 1$ (more measurement noise). In this simulation we assumed that $C$ and $D$ are known and $A$ and $B$ are to be estimated using (5.3.2) and (5.3.4). Our random initial guess is $\hat{A}_0 = \begin{bmatrix} 0.0346 & 0.5297 \\ 0.0535 & 0.6711 \end{bmatrix}$ which has eigenvalues of $1.3321$ and $0.4335$. The model estimate $\hat{\lambda}_k$ used in the $J_k$, $O_k$ filters is updated at each time instant by new parameter estimates $\hat{A}_{k|k, \hat{\lambda}_k}$. The estimated value of $A$ after 20000 points is $\hat{A} = \begin{bmatrix} -0.0060 & -0.7949 \\ 0.9772 & -0.0810 \end{bmatrix}$. Figure 5.2 shows convergence of $\hat{a}_{12}$. The system is transformed into companion form at each time step.

**Ex 3:** (Real Data) To examine the performance of the algorithm with more realistic data we identified the transfer system from torque measurement (input) and arm acceleration (output) in a mechanical robot system. The data is provided by the DAISY project [61]. The input signal is roughly zero mean Gaussian. We compare our identification scheme with the system identified using both input and output measurements and matlab’s *n4sid*. Matlab’s *n4sid* suggests $N = 8$ as the most likely model order. For simplicity, we consider only estimation of $A$. Using the $C_{n4sid}$ and $D_{n4sid}$ information (ie. the $C$ estimated by *n4sid* and the estimation input variance respectively) our algorithm was able to estimate the $A_{n4sid}$ from a variety of initializations without knowledge of the input signal (for our purposes here it is considered an unknown white noise input). We tried this for a variety of model orders. In particular for $N = 2$, $A_{n4sid} = \begin{bmatrix} 0 & -0.9788 \\ 1.0000 & 1.4007 \end{bmatrix}$ and $C_{n4sid} = [-0.1821, -0.2484]^T$, our parameter estimate is $\hat{A} = \begin{bmatrix} 0 & -0.9852 \\ 1.0000 & 1.4162 \end{bmatrix}$. This data set is relatively short (1024 data points) and several passes through the data is required.

### 5.6 Conclusions

Global convergence results have been developed for new finite dimensional adaptive schemes for estimation of the parameters of a partially observed discrete-time linear stochastic system and hybrid linear systems. The convergence results are developed using standard martingale properties and convergence results, the Kronecker lemma and an ordinary differential equation
approach. We emphasize that, for stable linear systems or hybrid linear systems driven by white noise, there is consistent estimation and, with the model estimate used in adaptive state estimation, this is asymptotically optimal.
Figure 5.1: Estimation of $a_{21}$

Figure 5.2: Estimation of $a_{12}$ in higher noise
Chapter 6

Optimal HMM Filtering and Differential Decoding

6.1 Introduction

A frequently occurring problem encountered in wireless digital communication systems is signal fading which results from multiple propagation paths [67]. Two complementary techniques to enhance detection of a message sequence transmitted over unknown channels are decision feedback and differential encoding schemes.

In message estimation, channel knowledge can be used to counter the distortions introduced by the channel, yet the transmission channel is frequently unknown, see [30, 49, 15, 14]. In 1965, Lucky [54] proposed an adaptive method for estimating the channel distortions now known as decision feedback. Decision feedback uses 'best' estimates of the transmitted message to estimate the transmission channel and hence improve message detection.

In slowly varying channel environments, message estimates from a Viterbi algorithm, proposed in 1967 by Viterbi[78] and invariably delayed to give improved estimation, can be used in a decision feedback structure [67, Page 651]. However, in wireless transmission the fading environment typically varies relatively rapidly [74] and channel estimates are required immediately. Decision feedback that is based on immediate message estimates is one technique for overcoming the type of channel distortion introduced in rapidly fading environments.

In differentially encoded signals, the message is encoded in the difference between successive symbols, see Proakis[67, Page 187] for further explanation. This means that the signals can be detected even when the transmission channel is unknown as long as the channel variation between successive symbols is small. Demodulation of differentially encoded signals is
therefore not contingent on explicit knowledge of the transmission channel. An important property of differential demodulation is that there is rapid recovery after channel nulls such as those that are typical in Rayleigh fading transmission channels. However, the performance of differential modulation can be poor when there is rapid channel variation.

Standard demodulation of differentially encoded signals should be viewed as a form of digital demodulation (or decision feedback) where the channel estimate is derived only from the previous symbol [67]. Here we generalize demodulation of differential encoded signals by using more information than just the last received symbol.

For clarity, in this chapter differential modulation, without trellis coding or bit interleaving, with a simple decision feedback structure is considered. The key contribution of this chapter is the proposal of the optimal HMM filter which exploits the idempotent property of Markov chains. Following this, a Kalman filter which also exploits the idempotent property is proposed together with a decision feedback structure to improve receiver performance. Compare this to the approach taken by Collings and Moore in [15] where a suboptimal HMM filter and a standard Kalman filter are used to perform decision feedback.

This chapter begins by introducing an indicator vector state space formulation of differential signalling. Using this state space formulation and by exploiting the idempotent nature of Markov states, an $O(N^3)$ optimal HMM filter is presented in an informative way which highlights the structure of the problem.

We next incorporate our HMM filter into a decision feedback structure. We propose a receiver structure where a conditional HMM filter is coupled to a conditional Kalman filter to incorporate decision feedback to aid demodulation. The standard differential demodulation scheme is presented to highlight the effect of rapidly changing channel conditions on its performance. Following this, the use of decision feedback on differentially coded transmission systems is investigated.

This chapter is organized as follows: In Section 6.2, we formulate the HMM signal model and channel model for a differential encoded system. In Section 6.3, the optimal HMM filter is presented. In Section 6.4, we present a standard differential receiver and highlight the effect of rapid channel variation on receiver performance. In Section 6.5, decision feedback based on a Kalman filter is proposed. In Section 6.6, simulation studies are presented. Finally, some conclusions are presented in Section 6.7.
6.2 State Space Formulation

In this section a model for differential phase modulation, and a model for the transmission channel are presented and then these are reformulated as an HMM and a state space stochastic model respectfully.

To simplify the discussion in this chapter, it is assumed that differential phase shift keying (DPSK) is used to transmit the signal. DPSK is a common type of differential modulation. Other forms of modulation not considered here that could be handled using this approach include: pulse amplitude modulation (PAM), quadrature amplitude modulation (QAM) and others.

It is assumed that the relevant match or correlation filter demodulators have been implemented and symbol synchronization and timing issues have been resolved. We proceed now with a signal space analysis.

6.2.1 Signal Model

In DPSK transmission schemes the carrier signal is transmitted as phase information over the channel. Let \( f_k \) be our message signal, being a real-valued discrete-time signal, where

\[
f_k \in Z_f = \{ Z_f^{(1)}, \ldots, Z_f^{(N)} \} \quad \text{where} \quad Z_f^{(i)} = (i/N)2\pi \in \mathbb{R}
\]  

and we denote the vector \( z_f \) as follows

\[
z_f = (Z_f^{(1)}, \ldots, Z_f^{(N)}).
\]  

In differential transmission systems the carrier symbol is the modulo sum of the message sequence. Let \( \theta_k \) denote the carrier symbol, then

\[
\theta_k = (\theta_{k-1} + f_k)_{2\pi}
\]  

where \((.)_{2\pi}\) denotes a modulo 2\(\pi\) operation.

The transmitted symbol at time \( k \), represented in the customary complex baseband notation,

\[
m_k = \exp(j\theta_k) \in \mathbb{C}
\]  

where imaginary and real components are transmitted using the quadrature and in-phase components of a carrier waveform. For simplicity unity amplitude is assumed.
6.2.2 Channel Model

The baseband signal $m_k$ is transmitted via a channel which can cause both amplitude attenuation and phase shift. The channel can be represented as a multiplicative disturbance, $g_k$, that is,

$$g_k = \kappa_k \exp(j\phi_k) = g_k^R + jg_k^I \in \mathbb{C} \quad (6.2.5)$$

where the superscripts $R$ and $I$ refer to the real and imaginary parts. This disturbance introduces time-varying gain and phase changes to the signal and is assumed to vary slowly.

The baseband observation process $y_k$ is thus assumed to have the form

$$y_k = g_k m_k + w_k \in \mathbb{C} \quad (6.2.6)$$

We define $Y_k := (y_0, ..., y_k)$ and $w_k \sim N(0, R_k)$. We assume $w_k$ is complex with real and imaginary parts that are i.i.d., with zero mean and Gaussian density, i.e. $w_k^R \sim N(0, \sigma^2_R)$ and $w_k^I \sim N(0, \sigma^2_I)$, where $w_k^R$ and $w_k^I$ are the real part and imaginary parts of $w_k$ respectively. Let $\mathcal{Y}_k$ denote the complete filtration generated by $y_\ell$, $\ell \leq k$. As a consequence,

$$E[w_{k+1} | \mathcal{Y}_k] = 0. \quad (6.2.7)$$

This model of the channel is simplistic in that it allows no inter-symbol interference (ISI) and assumes Gaussian noise, but it is realistic in narrow band communications. This channel model can represent fading channels through variations in $\phi_k$ and $\kappa_k$.

6.2.3 State Space Signal Model

A discrete-time state space model for the signal model in the previous section is now presented. Consider the following assumption on the message sequence, $f_k$

Assumption on the message source

$$f_k \text{ is a first order Markov chain} \quad (6.2.8)$$

For linear modulation without memory (such as DPSK and QAM) this assumption appears inappropriate because the symbols from the message source are usually assumed to be mutually independent. There would seem no advantage in viewing the message sequence as arising from a Markov chain. However, for various other modulation techniques such as NRZI
and Miller coding this Markov assumption appears more natural. This assumption also appears appropriate for the case when message symbols are convolutionally coded. The HMM filter presented below can be used to generate preliminary state estimates for the sole purpose of estimating the channel in systems with trellis or turbo coded signals with interleaving. The Viterbi algorithm could then later be used in parallel to produce the final message estimates.

Let us define an indicator vector \( X_k^f \in \{ e_1, ..., e_n \} \) associated with each message symbol, \( f_k \in Z_f \), where \( e_i = (0, ..., 0, 1, 0, ..., 0) \) with 1 in the \( i \)th position. That is, to each possible message symbol, \( Z_f^{(i)} \), we associate an indicator vector, \( e_i \). We can now write \( f_k \) in terms of \( X_k^f \) as

\[
f_k = z_f^f X_k^f.
\]  

Hence, under assumption (6.2.8) the transition probability matrix of the message process is

\[
A = (a_{ij}) \quad 1 \leq i, j \leq N
\]

where

\[
a_{ij} := P(X_{k+1}^f = e_i | X_k^f = e_j)
\]

so that

\[
E[X_{k+1}^f | X_k^f] = AX_k^f
\]

where \( E[\cdot] \) denotes the expectation operator. We also denote \( \{ \mathcal{F}_\ell, \ell \in \mathbb{Z}^+ \} \) the complete filtration generated by \( X_k^f \), that is, for any \( k \in \mathbb{Z}^+ \), \( \mathcal{F}_k \) is the complete filtration generated by \( X_k^f \), \( \ell \leq k \).

**Lemma 6.2.1** Under assumption (6.2.8) the dynamics of \( X_k^f \) are given by the state equation

\[
X_{k+1}^f = AX_k^f + M_{k+1}
\]

where \( M_{k+1} \) is a \((A, \mathcal{F}_k)\) martingale increment, in that \( E[M_{k+1} | \mathcal{F}_k] = 0 \).

**Proof** See [18]

Likewise, let us define an indicator vector, \( X_k^\theta \in \{ e_1, ..., e_n \} \), associated with carrier symbol, \( \theta_k \), such that

\[
\theta_k = z_f^\theta X_k^\theta
\]
Lemma 6.2.2 The indicator vectors $X_k^\theta$ and $X_k^f$ are related as follows

$$X_{k+1}^\theta = D(X_{k+1}^f)X_k^\theta \quad \text{or} \quad (6.2.15)$$

$$X_{k+1}^\theta = D(X_k^\theta)X_{k+1}^f \quad (6.2.16)$$

where $D(e_i) = S^{(e_i)}$ the shift operator. Note, $S$ is defined as

$$S = \begin{pmatrix}
0 & 0 & \cdots & 0 & 1 \\
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0
\end{pmatrix} \quad (6.2.17)$$

and $\mathbb{n} = (1, \ldots, N)'$.

Proof Equation (6.2.15) comes from equation (6.2.3). Equation (6.2.16) follows from noting that $X_{k+1}^f$ and $X_k^\theta$ are indicator vectors and as such non-linear functions can be written as linear functions, see [18].

The observation process can now be expressed in terms of the indicator vectors as follows.

$$y_k = \begin{pmatrix} y_k^R \\ y_k^I \end{pmatrix} = \begin{pmatrix} m_k^R & -m_k^I \\ m_k^I & m_k^R \end{pmatrix} \begin{pmatrix} g_k^R \\ g_k^I \end{pmatrix} + \begin{pmatrix} w_k^R \\ w_k^I \end{pmatrix} = \begin{pmatrix} (z_k^R)^T X_k^\theta \\ (z_k^I)^T X_k^\theta \end{pmatrix} \begin{pmatrix} g_k^R \\ g_k^I \end{pmatrix} + \begin{pmatrix} w_k^R \\ w_k^I \end{pmatrix} = H(z_k, X_k^\theta) \begin{pmatrix} g_k^R \\ g_k^I \end{pmatrix} + w_k$$

with appropriate definition of $H(z_k, X_k^\theta)$ and where the superscripts $R$ and $I$ refer to the real and imaginary parts.

Define two vectors of parametrized probability densities: $B_k = diag(b_k(e_1), \ldots, b_k(e_N))$, where $b_k(e_i) = P(Y_k | X_k^\theta = e_i)$ and $B_k|_i = diag(b_k|_i(e_1), \ldots, b_k|_i(e_N))$, where $b_k|_i(e_j) = P(Y_k | X_k^f = e_j, X_{k-1}^\theta = e_i)$.

In the special case, as here, when $w_k$ is complex and its components have Gaussian densities, we can write

$$b_k = \frac{1}{2\pi \sigma_R \sigma_I} \exp \left( - \frac{(y_k^R - H(z_k^R, e_i)^R g_k)^2}{2\sigma_R^2} - \frac{(y_k^I - H(z_k^I, e_i)^I g_k)^2}{2\sigma_I^2} \right) \quad (6.2.18)$$

and

$$b_{k|_i}(e_j) = a_{ji} b_{k-1}(e_i) \quad (6.2.19)$$

where $H(z_k^R)$ and $H(z_k^I)$ are the real and imaginary parts of $H(z_k)$ respectively.
6.3 Optimal HMM filter

Standard demodulation of DPSK is performed by phase comparison. Optimal path demodulation of convolutionally coded signals is performed by the Viterbi algorithm. Here we present the optimal HMM filter for message demodulation in two ways.

The key advantage of the HMM filter over standard demodulation techniques is that it provides instantaneous soft decision information. Standard demodulation techniques result in either maximum a posteriori estimates, from the matched filter-phase comparison, or delayed maximum likelihood estimates via the Viterbi algorithm.

The hidden Markov signal model offers the flexibility of being able to model or approximate communication systems with various degrees of complexity. For example, the HMM filter can be based on knowledge of only the digital modulation layer of a communication system, as we consider in this chapter, or include knowledge of the channel encoder layer for examples where the information source has been convolutionally coded. Due to the soft information that the HMM filter provides it also appears well suited to communication systems in which bit interleaving is performed. The HMM filter structure presented in this chapter can be used to generate preliminary soft message estimation for the purpose of channel estimation and the Viterbi algorithm, working in parallel and conditioned on these channel estimates, could be used to produce the hard decision on message estimates.

6.3.1 State Space Representation

Let $X^0_k$ denote the space of the modulation scheme. This space is also represented by the indicator vectors $X^f_k$ and $X^\theta_k$. The approach taken in the previous formulation of HMM filters is to obtain $X^0_k$ from the Kronecker product of these indicator vectors. That is,

$$X^0_k := X^f_k \otimes X^\theta_{k-1} \tag{6.3.1}$$

where $\otimes$ is the Kronecker product. $X^0_k$ is known to be a Markov process and standard HMM filtering theory can be applied. However, $X^0_k$ is $(N^2 \times 1)$ and hence the filter calculations are of order $N^4$, including zero operations.

If instead we define an indicator matrix $X_k$ as follows

$$X_k := X^f_k X^\theta_{k-1} \tag{6.3.2}$$


then we note that

\[ X_k^f = \mathcal{X}_k \mathbf{1}_N \quad \text{and} \quad X_{k-1}^g = \mathcal{X}_k^f \mathbf{1}_N \] (6.3.3)

where \( \mathbf{1}_N = (1, \ldots, 1)' \), an \( N \)-vector of ones.

**Lemma 6.3.1** The dynamics of \( \mathcal{X}_k \) are given by the state equation

\[ \mathcal{X}_{k+1} = \sum_{i=1}^{N} [e_i \mathcal{A}_i(A)] [e_i'^t \mathcal{X}_k]^t + \mathcal{M}_{k+1} \] (6.3.4)

where \( e_i'^t \mathcal{X}_k \) gives the \( i \)th row of matrix \( \mathcal{X}_k, \mathcal{A}_i(A) \) is a transition matrix, and \( \mathcal{M}_{k+1} = M_{k+1} e_i'^{(i \bmod N)} \) and is a \((A, \mathcal{F}_k)\) martingale increment, in that \( E[\mathcal{M}_{k+1} | \mathcal{F}_k] = 0 \).

**Proof**

\[ \mathcal{X}_{k+1} = X_{k+1}^f X_{k+1}^g \]

\[ = [AX_k^f + M_{k+1}]D(X_{k}^f)X_{k-1}^g \]

\[ = A \mathcal{X}_k^f X_{k-1}^g D(X_{k}^f) + M_{k+1}D(X_{k}^f)X_{k-1}^g \].

Now by noting that \( \mathcal{X}_k \) is a function of \( X_k^f \) and \( X_{k-1}^g \), that is, \( \mathcal{X}_k(X_k^f, X_{k-1}^g) = X_k^f X_{k-1}^g \) and denoting \( \mathcal{M}_{k+1} = M_{k+1}D(X_{k}^f)X_{k-1}^g \) we obtain

\[ \mathcal{X}_{k+1} = A \mathcal{X}_k(X_k^f, X_{k-1}^g)D(X_{k}^f) + \mathcal{M}_{k+1} \]

\[ = \left( \sum_{i=1}^{N} \sum_{j=1}^{N} A \mathcal{X}_k(e_j, e_i)D(X_{k}^f) \right) + \mathcal{M}_{k+1} \].

Using the idempotent property of indicator vectors we can write

\[ \mathcal{X}_{k+1} = \left( \sum_{i=1}^{N} \sum_{j=1}^{N} Ae_j e_i'^t D(e_j) \mathcal{X}_k^{(j,i)} \right) + \mathcal{M}_{k+1} \]

where \( \mathcal{X}_k^{(j,i)} \) denotes the \( j \)th element of \( \mathcal{X}_k \). Performing the inner summation, introducing \( \mathcal{A}_i \) and writing as a matrix product we obtain

\[ \mathcal{X}_{k+1} = \left( \sum_{i=1}^{N} \mathcal{A}_i e_i'^t \mathcal{X}_k \right) + \mathcal{M}_{k+1} \]

where

\[ \mathcal{A}_i = \begin{bmatrix} Ae_1 e_1'^t D(e_1) \\ Ae_2 e_1'^t D(e_2) \\ \vdots \\ Ae_N e_1'^t D(e_N) \end{bmatrix} \].
It follows from Lemma 6.2.2 that $\mathcal{M}_{k+1}^{(i,j)} = M_{k+1} e_{(i+j \mod N)}$. From Lemma 6.2.1, $M_{k+1}$ is a $(A,F_k)$ martingale increment and hence $\mathcal{M}_{k+1}$ is a $(A,F_k)$ martingale increment.

□

From Lemma 6.3.1 it follows that the dynamics of $\mathcal{X}_k$ can be viewed as $N$ parallel independent HMMs. That is, if we denote the $i$th row of $\mathcal{X}_{k+1}$ by $\mathcal{X}_{k+1}^{(i,j)}$, then,

$$\mathcal{X}_{k+1}^{(i,j)} = A_i \mathcal{X}_{k}^{(i,j)}.$$  

We can use standard HMM filter techniques on each of the rows of $\mathcal{X}$ to obtain an estimate of $\hat{\mathcal{X}}_k = E[\mathcal{X}_k | Y_k]$. That is, for each row,

$$\hat{\mathcal{X}}_{k+1}^{(i,j)} = N_k B_{k+1} A_i \hat{\mathcal{X}}_{k}^{(i,j)},$$  

(6.3.5)

where $N_k = \langle B_{k+1} A_i \hat{\mathcal{X}}_{k}^{(i,j)}, 1 \rangle^{-1}$ where 1 is the vector of ones.

Then $\hat{\mathcal{X}}_{k+1}^f$ can be obtained from $\hat{\mathcal{X}}_{k+1}$ using property (6.3.3).

Remarks

1. Lemma 6.3.1 states that the rows of $\mathcal{X}_k$ can be considered as Markov states that evolve to form rows of $\mathcal{X}_{k+1}$. This structure is hidden in the $\mathcal{X}_0^0$ formulation.

2. This filter requires $O(N^3)$ calculations per time instant.

6.3.2 Conditional Filter Formulation

In this subsection a more convenient formulation of our optimal HMM filter using coupled conditional HMM filters is presented. In this formulation conditional HMM filters are introduced for $X_{k+1}^f$ and $X_{k+1}^\theta$ which exploits the interdependence between signals $\{f_k\}$ and $\{\theta_k\}$.

Let $\hat{X}_{k+1}^f$ and $\hat{X}_{k+1}^\theta$ denote the normalized state estimates of $X_{k+1}^f$ and $X_{k+1}^\theta$ obtained from their respective conditional filters. That is,

$$\hat{X}_{k+1}^f := E[X_{k+1}^f | Y_k],$$  

$$\hat{X}_{k+1}^\theta := E[X_{k+1}^\theta | Y_k].$$  

(6.3.6)

First, consider the intermediate conditional state estimate, $\hat{X}_{k+1}^f$, given by

$$\hat{X}_{k+1}^f := E[X_{k+1}^f | Y_k, X_k^\theta = e_i].$$  

(6.3.7)
From Bayes’ rule it is clear that

$$\hat{X}_{k+1}^f = \sum_{i=1}^{N} \hat{X}_{k+1|i}^f \hat{X}_k^f(i) \quad (6.3.8)$$

where $\hat{X}_k^f(i)$ is the $i$th element of $\hat{X}_k^f$. Note that $\hat{X}_k^f(i) = P(X_k^f = e_i | Y_k)$.

**Lemma 6.3.2** The following forward recursion exists to estimate $\hat{X}_{k+1|i}$

$$\hat{X}_{k+1|i}^f = N_{k}^{(1)} B_{k+1|i} A \hat{X}_k^f \quad (6.3.9)$$

where $N_{k}^{(1)} = (B_{k+1|i} A \hat{X}_k^f, \mathcal{U})^{-1}$ is a normalizing factor.

**Proof** Follows from assumption (6.2.8), Lemma 6.2.1 and standard HMM theory. \qed

**Lemma 6.3.3** The conditional filtered normalized state estimate $\hat{X}_{k+1}^\theta$ is obtained by

$$\hat{X}_{k+1}^\theta = N_{k}^{(2)} B_{k+1} \sum_{i=1}^{N} D(e_i) \hat{X}_{k+1|i}^f \quad (6.3.10)$$

where $N_{k}^{(2)} = (B_{k+1} \sum_{i=1}^{N} D(e_i) \hat{X}_{k+1|i}^f, \mathcal{U})^{-1}$ is a normalizing factor.

**Proof** Follows from (6.2.16) and Bayes’ rule. \qed

**Lemma 6.3.4** The conditional filtered normalized state estimates $\hat{X}_{k+1}^f$ is given by

$$\hat{X}_{k+1}^f = N_{k}^{(3)} \sum_{i=1}^{N} \hat{X}_{k+1|i}^f e_i \hat{X}_{k+1}^\theta \quad (6.3.11)$$

where $N_{k}^{(3)} = (\sum_{i=1}^{N} \hat{X}_{k+1|i}^f e_i \hat{X}_{k+1}^\theta, \mathcal{U})^{-1}$ is a normalizing factor.

**Proof** Follows from definition of $\hat{X}_{k+1}^f$ and (6.3.8) \qed

Application of these last three lemmas gives a recursive filter for estimating $\hat{X}_{k+1}^f$, and hence $f_{k+1}$, at each time instant.

**Remarks**

1. The primary difference between the suboptimal approach in [12] and the formulation here appears in Lemma 6.3.3. The suboptimal approach would update $\hat{X}_{k+1}^\theta$ as follows

$$\hat{X}_{k+1}^\theta = B_{k+1} \hat{X}_{k}^\theta$$

rather than (6.3.10).

2. These filters require $O(N^3)$ calculations per time step to implement.
6.4 Differential Receiver

In this section the standard differential receiver is introduced to highlight the effect of channel variations on the performance of the standard receiver.

In the standard receiver the transmitted symbol is recovered from the quadrature and in-phase components of the carrier waveform. Then, for the case of DPSK, the message symbol is recovered from the difference between successive received symbols. Let \( \theta_k \) denote the received phase at time \( k \). The received phase \( \tilde{\theta}_k \) is related to the transmitted phase \( \theta_k \) as follows, 

\[
\tilde{\theta}_k = (\theta_k + \theta_k^C + \phi(w_k))_{2\pi}
\]  

(6.4.1)

where \( \theta_k^C \) is the phase of the channel at time \( k \) and \( \phi(w_k) \) is the phase of \( w_k \). The difference between successive received phase signals is 

\[
(\tilde{\theta}_k - \tilde{\theta}_{k-1})_{2\pi} = (f_k + \Delta_k^C + \phi(w_k - w_{k-1}))_{2\pi}
\]  

(6.4.2)

where \( \Delta_k^C = \theta_k^C - \theta_{k-1}^C \).

When no information is available about the channel phase the standard estimate of the message symbol, see Figure 6.1, is 

\[
\hat{f}_k = (\tilde{\theta}_k - \tilde{\theta}_{k-1})_{2\pi}
\]  

(6.4.3)

It is clear from (6.4.2) that the quality of the estimate (6.4.3) will degrade as the rate of change of the channel, or \( \Delta_k^C \), increases.

However, if estimates of the channel phase, \( \hat{\theta}_k^C \), are available then these can be used to improve the performance of the receiver by including these estimates into the receiver

![Figure 6.1: Block diagram of M-ary differential coherent receiver](image-url)
structure. Consider the message estimate \( \hat{f}_k \) which includes channel estimates,

\[
\hat{f}_k = (\hat{\theta}_k - \hat{\theta}_{k-1} - \hat{\Delta}_k^{c})_{2\pi}
\]

(6.4.4)

where \( \hat{\Delta}_k^{c} = \hat{\theta}_k - \hat{\theta}_{k-1} \), see also Figure 6.2.

![Figure 6.2: Block diagram of differential receiver using channel estimates](image)

In this receiver the message estimate are related to the message as follows:

\[
\hat{f}_k = (f_k + \hat{\Delta}_k^{c} - \hat{\Delta}_k^{c} + \phi(w_k - w_{k-1}))_{2\pi}
\]

(6.4.5)

comparing with (6.4.2) it is clear that the performance of the receiver that uses the channel estimates will be better when \( |\Delta_k^{c} - \hat{\Delta}_k^{c}| < |\Delta_k^{c}| \).

This is the motivation for the following section.

### 6.5 Decision Feedback Channel Estimation

To produce channel estimates we propose the use of a coupled conditional hidden Markov model (HMM) filter and conditional Kalman filter (KF).

More complicated decision feedback structures, using banks of Kalman filters conditioned according to the soft estimate information (as suggested in [30]) or more complicated structures, are plausible but not considered here.

#### 6.5.1 Conditional HMM filter

Let \( \hat{X}_{k|c_k} \) denote the state estimate of \( X_k \) conditioned on channel information. That is,

\[
\hat{X}_{k|c_k} = E[X_k|Y_k,c_k]
\]

(6.5.1)

where \( c_k = \{ \bar{g}_0, \ldots, \bar{g}_k \} \).
Lemma 6.5.1 The estimates $\hat{X}_{k|c_h}$ can be found using the forward recursion

$$
\hat{X}_{k+1|c_h} = \hat{X}_k + \sum_{i=k}^{k+1} \mathcal{A}_i \hat{X}_{i|c_h}
$$

(6.5.2)

where $N_k = (B_{k+1|c_h} A_i \mathcal{A}_i^{-1} A_i B_{k+1|c_h})^{-1}$ is a normalizing factor for each row and $B_{k+1|c_h} = \text{diag}(b_{k+1|c_h}(e_1), \ldots, b_{k+1|c_h}(e_N))$, and $b_{k+1|c_h}(e_j) = P[Y_k|X_{k+1} = e_j, X_k = e_i, g_k = g_h],

\mathcal{A}_i = \begin{bmatrix}
Ae_1 e_i D(e_1) \\
Ae_2 e_i D(e_2) \\
\vdots \\
Ae_N e_i D(e_N)
\end{bmatrix}

and $\mathcal{M}_{k+1} = M_{k+1} e_{(i+j \text{ mod } N)}$.

Proof See Lemma (6.3.1).

6.5.2 Conditional KF Channel Estimate

The observation process (6.2.6) is bi-linear in the channel parameter, $g_k$, and the message symbol, $m_k$. In this subsection a conditional KF is proposed for estimation of the channel parameter given the message symbol.

If we assume the channel dynamics are given by the following linear time invariant stochastic system

$$
g_{k+1} = Fg_k + v_k \quad \in \mathbb{C}
$$

$$
y_k = H(z_f, X_k^0) g_k + w_k \quad \in \mathbb{C}
$$

(6.5.2)

where $F \in \mathbb{R}^{2 \times 2}$, $H(z_f, X_k^0) \in \mathbb{R}^{2 \times 2}$, $v_k = N(0, Q_h)$ and $w_k = N(0, R_h)$. Then the KF equation for estimating $g_k$ given the message estimate $X_k^0$ is,

$$
\hat{g}_k = F\hat{g}_{k-1} + K_k[g_k - \hat{g}_{k-1} m_k]
$$

$$
K_k = \Sigma_{k|k-1} R_h^{-1} \text{ diag}(H_k) \Sigma_{k|k-1} R_h^{-1}
$$

$$
\Sigma_{k|k} = (I - K_k H_k^\prime ) \Sigma_{k|k-1}
$$

$$
\Sigma_{k+1|k} = F \Sigma_{k|k} F^\prime + Q_h
$$

(6.5.3)

where we have written $H_k := H(z_f, X_k^0)$. We note that

$$
\Sigma_{k|k}^{-1} = \Sigma_{k|k-1}^{-1} + H(z_f, X_k^0) R_h^{-1} H(z_f, X_k^0)^\prime
$$

$$
= \Sigma_{k|k-1}^{-1} + \begin{bmatrix}
\sigma_{L}^2 L_k(X_k^0) & 0 \\
0 & \sigma_{w}^2 L_k(X_k^0)
\end{bmatrix}
$$

(6.5.4)
where

\[ L_k(X_k^\theta) = (z_f^R)'X_k^{\theta_f}z_f^R + (z_f^I)'X_k^{\theta_I}z_f^I \]

\[ = (z_f^R)' \text{diag}(X_k^\theta) z_f^R + (z_f^I)' \text{diag}(X_k^\theta) z_f^I, \quad (6.5.5) \]

and where \( \text{diag}(X_k^\theta) \) is the diagonal matrix with \( X_k^\theta \) on its diagonal.

In (6.5.5) we have exploited the idempotent nature of Markov chains by replacing the product \( X_k^\theta X_k^{\theta_f} \) by \( \text{diag}(X_k^\theta) \). The effect of exploiting this idempotent property here is to force the Kalman filter to incorporate this \textit{a priori} knowledge. In the following section \( X_k^\theta \) is replaced by the conditional mean estimate \( \hat{X}_k^\theta \), however \( \hat{X}_k^\theta \hat{X}_k^{\theta_f} \neq \text{diag}(\hat{X}_k^\theta) \). Other formulations of the Kalman filter for this problem, for example [15] and [67, Page 656], do not exploit this structure.

To reduce computational effort we note that if \( \Sigma_{\theta_0}^{-1} \), \( F \) and \( Q_k \) are diagonal for all \( k \) then \( \Sigma_{k|k}^{-1} \) and \( \Sigma_{k+1|k}^{-1} \) will be diagonal and the Kalman filter equations can be simplified as.

\[
\hat{g}_k = F\hat{g}_{k-1} + K_k[y_k - \hat{g}_{k-1}m_k]
\]

\[
K_k = F\Sigma_{k|k}^{-1}\bar{R}_k R_k^{-1}
\]

\[
\Sigma_{k|k}^{-1} = \Sigma_{k-1|k-1}^{-1} + \begin{bmatrix}
\sigma_w^R L_k(X_k^\theta) & 0 \\
0 & \sigma_w^I L_k(X_k^\theta)
\end{bmatrix}
\]

\[
\Sigma_{k+1|k} = F\Sigma_{k|k} F' + Q_k. \quad (6.5.6)
\]

This simplification for the \( K_k \) recursion follows from the matrix inversion lemma in the same way as [1, Page 140]

Estimates for the channel phase, \( \hat{\theta}_k^c \), are given by \( \hat{\theta}_k^c = \phi(\hat{g}_k) \).

\textbf{Remarks}

1. Often, even when the channel dynamics are not exactly in the form (6.5.2) they can be approximated by (6.5.2).

2. \( F \) is typically approximated to be of the form \( fI \) for some \( f < 1 \) which allows equations (6.5.6) to be used and reduces computational effort. A forgetting factor can be introduced into line 4 of recursions (6.5.6) to allow for model variations and modelling errors.

3. The Kalman filter requires \( O(N^2) \) calculations per time instant. The simplified Kalman filter requires \( O(N) \) calculations per time instant.
6.5.3 The Complete Algorithm

To allow simultaneous estimation of the channel, $\hat{y}_k$, and the message, $\hat{m}_k$ or $\hat{X}_k$, the conditional filters (6.5.2) and (6.5.3) are coupled together, see Figure 6.3.

![Diagram of coupled filter structure](image)

Figure 6.3: Block diagram of coupled filter structure

Remarks

- The coupled algorithm requires $O(N^3)$ calculations per iteration to implement.
- The noise variance in both the HMM filter and the KF filter may need to be increased to "model" the induced noise has been suggested in Chapter 2.
- The decision feedback structure does not suffer from the usual error propagation effects. While message estimation errors may indeed introduce constant phase errors in the channel phase estimate, these absolute phase errors do not degrade the performance of the differential receivers, as is well known [82]. In the usual way, a decoding error will be followed by at least one more decoding error.

6.6 Simulations

6.6.1 Differential Decoding With and Without Decision Feedback

In our simulations we investigate the gain in performing decision feedback on an uncoded transmission system. To evaluate the performance of the receiver in a likely environment the transmitted message sequence consists of an i.i.d. message sequence, convolutionally coded using a rate 2/3 convolutional code. The HMM-KF filter is used to estimate the carrier symbols not the message sequence. This environment represents the situation in which the
HMM-KF filter is used to produce channel estimates that are passed to a Viterbi algorithm to perform optimal decoding of the i.i.d. message sequence.

The encoded signal is transmitted over a transmission channel that is time varying and unknown. The variations are deterministic, with the phase and amplitude varying sinusoidally from 50% to 150% of a nominal value. The optimal HMM decoder is used to demodulate the encoded symbols, but the original i.i.d. message sequence is not decoded from the demodulated symbols. In a realistic application, the Viterbi algorithm would be used in parallel for optimal decoding.

Figure 6.4 shows an improvement in bit error performance in terms of the encoded symbols due to decision feedback. Likewise, the channel estimates are seen to improve the bit error rate in terms of the i.i.d. message sequence. This curve demonstrates that there is over half a dB gain to achieve $P_e = 10^{-3}$.

### 6.7 Conclusion

In this chapter optimal hidden Markov model (HMM) filtering and decision feedback of differentially encoded transmission signals has been investigated. We have presented the optimal HMM filter for demodulation of differentially encoded signals. We then proposed a decision feedback structure that coupled together a conditional HMM filter to a conditional Kalman filter, a key point being that the Kalman filter in exploiting the idempotent nature of Markov chains included more a priori structural information. We also presented simplified Kalman filters for particular assumptions in the channel. The HMM filter requires $O(N^3)$ calculations per time instant, where $N$ is the number of message symbols. The Kalman filter requires $O(N^2)$ calculations per time instant which is reduced to $O(N)$ under the new channel assumptions.

Simulation studies demonstrated a half dB gain in signal to noise ratio required to achieve a bit error rate of $1 \times 10^{-3}$ symbols/symbol.
Figure 6.4: Improvement in BER performance due to decision feedback.
Chapter 7

Concluding Remarks and Future Work

Who, proudly seized of learning’s throne,
Now damns all learning but his own.

From The Ghost by Charles Churchill

7.1 Introduction

This chapter presents a summary of the major contributions of the research presented in this thesis. Further research questions are also posed.

7.2 Summary of Contributions

New Estimation Algorithms

The new estimation algorithms contributed in this thesis were:

1. New asymptotically efficient parameter estimators for hidden Markov models using recursive prediction error (RPE) techniques, see Chapters 2 and 3.

2. New suboptimal parameter estimators for HMMs using extended least squares (ELS) techniques, see Chapters 2 and 3.

3. Almost surely convergent parameter estimators for hidden Markov models, see (4.4.1).
4. Almost surely convergent parameter estimators for partially observed linear and hybrid linear systems, see (5.3.4) and (5.4.23).

5. Optimal $O(N^3)$ filter for estimation of differentially encoded message sequences, see Chapter 6.

6. Suboptimal parameter estimators for HMM parameters using various approximations of almost sure estimators, see Section 4.4.3.

**New Theory/Techniques**

The major theoretical contributions of this thesis were:

1. A new cost index for identification of state values that uses filtered estimates rather than one-step-ahead predictions, see Chapter 2.

2. A new cost index for identification of transition probabilities, see Chapter 3.

3. Partial convergence results for the ELS algorithms using standard martingale convergence results and the Kronecker lemma, see Chapters 2 and 3.

4. The optimal HMM transition count and the optimal observation count filters in matrix form, see (4.3.15).

**New Application Algorithms**

The following application algorithms were proposed in this thesis:

1. An HMM-KF feedback structure that exploited the idempotent property of HMMs to improve demodulation.

Below is a more detailed description of the major contributions of this thesis.

**RPE Algorithms for Estimating the Output Mapping Vector of an HMM**

Chapter 2 examined recursive prediction error methods for estimating the state output mapping vector of an HMM from observations (assuming that the state transition probability matrix is known). In this chapter new on-line schemes based on extended least squares (ELS)
and recursive prediction error (RPE) methods were presented. Convergence result were established for the RPE algorithms using an ordinary differential equation (ODE) approach. These new schemes exploit the idempotent property of Markov chains to reduce the computational effort. The algorithms presented have computational complexity of $O(N)$ yet perform as well asymptotically as earlier schemes propose of $O(N^2)$.

Of all the schemes presented in Chapter 2 the *a posteriori* ELS and *a posteriori* weighted RPE scheme, which exploit filtered state estimates rather than prediction estimates, appear to be the most attractive for application purposes. These *a posteriori* schemes have been also found to be consistent and thus attractive in signalling environments that include low inertia HMMs which could not be handled well by earlier algorithms [13].

Implementation aspects were discussed including estimation of time-varying parameters, variance corrections, and estimation in extremely high noise. Illustrative simulation examples of the schemes in a variety of conditions were presented. These simulations highlighted the similarities and the differences between the schemes. The simulation studies also compared the new schemes with the RPE scheme of [13].

### RPE Algorithms for estimating the State Transition Probability Matrix of an HMM

In Chapter 3 we proposed new algorithms for recursive estimation of the state transition probability matrix of an HMM based on ELS and RPE techniques. The major contribution of this chapter was the proposal of new criteria for identification of the state transition probability matrix which avoids the ill-conditioning, in low noise, of the schemes in [13]. An explanation of why this criteria is more appropriate than the standard least squares cost is given.

Convergence analysis for the recursive (state) prediction error (RSPE) algorithm was provided via an ODE approach but no convergence results were presented for the ELS algorithm. Despite the lack of convergence results, the ELS algorithm is attractive because it has computational complexity of only $O(N^2)$ per time instant, compared with the RPE scheme (of [13]) and the RSPE scheme of this chapter which have computational complexity $O(N^4)$.

This chapter we also proposed a scheme for the simultaneous estimation of state output mapping values and the state transition probabilities. Local convergence results were presented. Illustrative simulation studies were presented that demonstrate the performance of
these algorithms.

**Almost Sure Parameter Estimation for HMMs**

Chapter 4 investigated almost sure parameter estimation for HMMs. The major contribution of this chapter was the proposal of almost surely convergent parameter estimators for HMMs. The online adaptive algorithm present in this chapter appears related to the off-line Baum-Welch EM algorithm; however, convergence results were established independently of EM algorithm's theoretical basis using convergence results from probability theory. Preliminary convergence results (when the state sequence is known or when idealized conditional mean estimates are available) were established using martingale convergence results and the Kronecker lemma. Complete global convergence results and convergence rates results (when the conditional mean estimates based on the true model are not available) were established using an ODE approach.

We emphasize that under reasonable excitation conditions, there is consistent estimation and estimation is asymptotically optimal.

Suboptimal algorithms that require reduced computational effort were proposed. Convergence results were not established for these suboptimal algorithms.

Simulation examples were presented.

**Almost Sure Parameter Estimation for Linear Systems and Hybrid Linear Systems**

The major contribution of Chapter 5 was the proposal of almost sure parameter estimators for partially observed linear systems and hybrid linear systems. Preliminary convergence results were established for linear systems and hybrid linear systems (in idealized situations when the state sequence is known or conditional mean estimates are available) using martingale convergence results and the Kronecker lemma.

Complete global convergence results and convergence rates results for estimation of linear systems, when the conditional mean estimates based on the true model are not available, were established using an ODE approach.

Simulation studies were presented for estimation of linear systems. The proposed parameter estimators for hybrid linear system were too computationally intensive and hence simulation studies were not given.
Demodulation of Differential Encoded Messages

Demodulation of differentially encoded messages using HMM filters was investigated in Chapter 6. The major contribution of this chapter was the proposal of a new state-space representation which lead to the proposal of the \( O(N^3) \) optimal HMM filter. Two methods of implementing the optimal HMM filter were given. The HMM filter provides immediate soft decision information which appears to be important for improved receiver design when bit interleaving or other encoding techniques are incorporated in the transmission system.

Decision feedback was investigated for differentially encoded sequences using the proposed optimal HMM filter and a KF channel equalizer. Simulation examples were presented and it was demonstrated that decision feedback improves receiver performance. The performance gains were appealing but more work needs to be done before this approach can be considered as a practical alternative.

7.3 Future Research Directions

In this thesis several open questions have arisen about identification of hidden Markov models.

1. If the “true model in the model set” assumption is relaxed then what do our model estimates mean? Will our estimates be maximum likelihood estimates? Can maximum likelihood estimation on a set of HMMs of a particular model order be performed? Consider estimation of an HMM of order \( N \) in the following situations:

   (a) the data is generated by an HMM of an order different than \( N \).

   (b) the data is generated by a linear system.

   (c) the data is generated by a non-linear system.

2. Is there a concept of a minimal representation? When will high-order HMMs be ‘well’ approximated by low-order HMMs?

3. Similar questions can be proposed about our estimators for linear systems. In what sense are model estimates ‘good’ when the model is not in the model set? etc.

Several issues remain unresolved about the work in Chapters 2 and 3. The following questions remain and may be worth investigating.
4. A recent publication suggests that the least mean squares algorithm is optimal in the 
\( H^\infty \) sense while the least squares algorithm is optimal in the \( H^2 \) sense [32]. The ap-
proximation introduced into the Hessian of our RPE type recursions in Chapters 2 and 3 is analog-
ous to the difference between least mean squares and least squares. Has this Hessian approxima-
tion made our recursions robust, in the sense of being robust to model uncertainty?

(a) This robustness issue may also be significant for the almost sure estimators we have 
proposed for HMMs, linear systems, and hybrid linear systems. Does the analogy 
apply to our almost sure estimators?

(b) In general, can \( H^2 \) optimality can be traded for robustness? For example, is it pos-
sible to modify linear system estimators to trade some \( H^2 \) optimality for robustness 
or vice versa?

5. This leads into a general question about robust or risk sensitive estimation. Can a 
risk sensitive (or robust) parameter estimation problem be posed? Risk sensitive filters 
appear to be appropriate because filtered estimates of quantities are required when the 
true model is not known but model estimates are available.

Here is a list of possible minor modifications to the almost sure HMM estimators for 
completeness.

6. It may be possible to modify the estimators of chapter 4 to allow the tracking of slow 
parameter variations. This could be done in an ad hoc way via forgetting factors (see 
Chapters 2 and 3).

7. To overcome the difficulty of estimating parameters in high noise it may be worth 
investigating the use of Polyak’s acceleration techniques or the use of forgetting factors 
(see Chapters 2 and 3).

8. The almost sure results could be extended to the discrete observation case.

More work in the area of demodulating differentially encoded signals would include simul-
ation of systems with more realistic transmission channels. The following questions/changes 
could be considered.
9. Does our algorithm offer improved bit error rate (BER) performance when the transmis-
sion occurs over more realistic channels? To make the channel more realistic, we should
simulate a Rayleigh fading environment. We suspect that in fast fading environments
our receiver may not offer direct BER performance improvement. This leads to the
following question.

10. When the soft decoding information is coupled/passed to a Viterbi algorithm decoder,
will our receiver improve BER performance? Many questions remain as to how to best
couple our receiver to a Viterbi algorithm decoder.

11. In slow fading environments, can our new global convergence parameter estimators be
used to track channel variations and hence improve BER performance? Consider the
following model:

\[ X_{k+1} = AX_k + V_k \]
\[ y_k = CX_k + w_k \]

where \( X_k \) is the message sequence, \( y_k \) is the received signal, \( C \) is the channel, and \( w_k \)
is additive channel noise. By introducing a forgetting factor into our estimator for \( C \)
we could track slow channel variations. Recently, there have been publications which
use the off-line EM algorithm in this way \[2\]. It may be worth investigating how much
improvement our on-line estimation algorithms offer.

12. The work in this thesis has assumed that the sampling rate of received symbols (chip
rate) is the same as the message bit rate. In environments where the chip rate is greater
than the bit rate, does our Kalman filter-HMM decoder receiver offer performance im-
provements? Our present simulation investigation suggests that the rate of channel
fading is too great for our Kalman filter to track. Multiple sampling of each message bit
will provide more information and may aid channel tracking performance of the Kalman
filter.

13. Presently, our simulation study assumes that the channel has a constant cut-off fre-
quency, an assumption that will not be valid in some transmission systems. However,
because our low order Kalman filter appears fairly insensitive to channel cut-off fre-
quency, it may be possible to relax this assumption. How does the performance of our receiver compare to the standard receiver in this environment?
Bibliography


153


[58] A.A. Markov, “Extension of the law of large numbers to dependent events,” (Russian) 


[61] De Moor B.L.R. (ed.), DAISY: Database for the Identification of Systems, De-
partment of Electrical Engineering, ESAT/SISTA, K.U.Leuven, Belgium, URL: 
http://www.esat.kuleuven.ac.be/sista/daisy/, 1/11/97. [Used dataset: Data from a 
flexible robot arm, Section: Mechanical Systems, Code: 9 6-009.]


tion,” *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. SMC-9, No. 4, pp 197-205, 
April, 1979.


[65] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averag-


