THESE

présentée pour obtenir le grade de

Docteur de l'Université de Nice Sophia-Antipolis (UNSA)

Spécialité: Automatique, Traitement du Signal et Informatique

Tayeb SADIKI

Filtrage Adaptatif Bayesien

La Thèse est soutenue le 02 Mars 2007, devant le jury composé de :

Mohamed Najim	Rapporteur
Ali Khenchaf	Rapporteur
Gerard Favier	Examinateur
Dirk Slock	Directeur de thèse

THESIS

presented to obtain the degree of Doctor of Philosophy

of the University of Nice Sophia-Antipolis (UNSA)

Specialization: Control, Signal Processing and Computer Science

Tayeb SADIKI

Bayesian Adaptive Filtering

Will be defended on March 2nd, 2007, before the committee composed by:

Mohamed Najim	Reader
Ali Khenchaf	Reader
Gerard Favier	Examiner
Dirk Slock	Thesis supervisor

To My parents

Notations

We regroup here the principal notations used in the different chapters of the thesis. As far as possible we tried to conserve the same notations from one chapter to another.

x	Scalar variable
X	Vector
$\{x_n\}$	Set of the variables x_n
x^T	Transpose of vector x
z^*	Complex conjugate of variable z
$\mid z \mid$	Magnitude of complex variable z
z^H	Hermitian, i.e., transpose conjugate, of complex vector z
A^{-1}	Inverse of matrix A
I	Identity matrix
0	All zero vector
$\mathbf{1}$	All one vector
$diag(x_0, \ldots, x_{N-1})$	Diagonal matrix with diagonal elements $\{x_n\}$
$\prod_n x_n$	Product of the N elements $\{x_n\}$
$\sum_n x_n$	Summation of the N elements $\{x_n\}$
$E\{x\}$	Expected value of random variable z

$var\{x\}$	Variance of random variable z
$Pr\{A\}$	Probability of event A to occur
p_x	Probability density function of random variable x
$PrA \mid B$	Probability of event A conditioned on event B
$\propto$	Proportional to
$\simeq$	Approximately equal to
$\bigotimes$	Convolutional product
$\bigodot$	Compound function, $f \bigodot g(x) = f(g(x))$
$argmin(f(x))$	Value of x that minimizes the function f(x)
$argmax(f(x))$	Value of x that maximizes the function f(x)
$\delta(x)\,, \delta_{ij}$	Dirac and Kronecker delta functions
$exp(x)$	Exponential function
$Q(x)$	Complementary error function
$J_0(x)$	Zero-order Bessel function of the first kind rect
$T(t)$	Rectangular function

Acronyms

Here are the main acronyms used in this document. The meaning of an acronym is usually indicated once, when it first occurs in the text. The english acronyms are also used for the french summary.

AWGN	Additive White Gaussian Noise
BAF	Bayesian Adaptive Filtering
BQUE	Best Quadratic Unbiased Estimator
CDF	Cumulative Distribution Function
DB	Doppler Bandwidth
DFT	Discrete Fourier Transform
DS	Direct-Spread
EM	Expectation Maximization
EMSE	Excess Mean Square Error
EKF	Extended Kalman Filter
FIR	Finite Impulse Response
FF	Forgetting Factor
IIR	Infinite Impulse Response
iff	if and only if
i.i.d.	independent and identically distributed
ISI	Inter-Symbol Interference
LSTBC	Linear Space-Time Block Code
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error

MSE	Mean Square Error
NLMS	Normalized Least Mean Square
LMS	Least Mean Square
OFDM	Orthogonal Frequency Division Multiplexing
RLS	Recursive least square
PDP	Power Delay Profile
RPEM	Recursive Prediction-Error Method
r.v.	Random Variable
Rx	Receiver
KF	Kalman Filter
SAF	Standard Adaptive Filtering
SBEM	Subspace-Based Estimation Method
SOS	Second-Order Statistics
SIMO	Single-Input Multiple-Output
SIR	Signal-to-Interference Ratio
SISO	Single-Input Single-Output
SNR	Signal-to-Noise Ratio
S/P	Serial-to-Parallel Conversion
s.t.	such that
SS	Step-size
STC	Space-Time Code
VSS	Variable Step-size
TV	Time-Varying
WF	Wiener Filter
w.r.t.	with respect to
WSS	Wide-sense Stationary

Acknowledgement

First and foremost I wish to thank my supervisor, Professor Dirk T. M. SLOCK, who gave me the opportunity to work on interesting problems at my own pace. His deep insight and experience in so many different areas was definitely a great help in understanding some of the finer points of digital communications. I am truly fortunate to have worked with him.

My friends in the Eurecom community have made my stay on the Cote d'Azur the most memorable experience of my life. Although I am leaving out many people, who I hope will not hold it against me, I wish to thank in particular David Gesbert. Eurecom is a truly wonderful place which I hope will continue to grow and prosper. My father's moral support was instrumental in my obtaining my doctorate. Everything I have accomplished is due to him. My late mother will always be in my heart and has always been able to guide me in her own special way. I wish to thank my grandmother who has always been a source of wisdom and encouragement. Finally, I must thank Cathy. Her unfailing love and strength was an enormous help during the final stage of my studies. Putting up with me while I was hospitalized and the following month at home was not an easy task. I can only hope to play the same role in her life as she does in mine.

Contents

List of Figures

Résumé

Le filtrage adaptatif est en principe destiné à poursuivre les systèmes mobiles. Cependant, la plupart des algorithmes de filtrage adaptatifs ont été conus pour converger à un filtre inconnu fixe. Une fois réellement confrontés avec un environnement mobile, ils possèdent juste un paramètre (stepsize, facteur d'oublie) pour ajuster leurs possibilités de poursuite. Dans le cas stationnaire de la non stationnarité, les coefficients optimaux de filtre évoluent comme un processus stationnaire. L'approche de filtrage adaptatif bayésien exploite l'information a priori dans ce modèle stationnaire de variation de paramètre pour optimiser l'exécution de filtrage adaptative. L'information préalable contient deux caractéristiques critiques de paramètre : la variance (l'amplitude) des divers coefficients de filtre et leur spectre de variation (le power delay profile (PDP) et le spectre de Doppler dans le cas du canal sans fil). Les outils pratiques pour mettre en application le filtrage adaptatif Bayésien (BAF) sont les filtres de Wiener ou de Kalman. Ce dernier modélise typiquement la variation des coefficients du filtre optimal comme des processus AR(1). Les paramètres de ces processus AR(1) peuvent être commodément identifiés avec un algorithme d'EM (Expectation Maximization) adaptatif. Pour limiter la complexité au même ordre que la complexité de l'algorithme de RLS, un modèle d'état diagonal pour Kalman peut être pris. Dans cette thèse on introduit ces techniques pour rendre le concept existant du filtrage adaptatif Bayésien praticable. Aussi, nous analysons l'effet du PDP et de la largeur de bande de Doppler en régime permanent des algorithmes Bayésiens et nous comparons avec les algorithmes standards (LMS et RLS). Dans la deuxième partie de cette thèse nous introduisons une approche à deux étapes; le but de cette approche est de présenter des techniques de BAF qui ne sont pas immensément plus complexes que l'algorithme LMS. L'approche à deux étages proposée se compose d'une première étape

utilisant un filtre adaptatif classique à convergence rapide, suivi d'un filtre passe-bas et de sous échantillonnage dans le temps des coefficients du filtre adaptatif. La deuxième étape applique alors un Kalman, filtrant à cadence réduite sur un modèle d'état simplifié. Finalement on aussi étudié l'optimisation de la fenêtre dans l'algorithme RLS, dans cette partie, nous considérons la poursuite d'un filtre optimal modélisé par un processus stationnaire vectoriel. Nous interprétons l'algorithme du moindre carré récursif (Recursive Least Squares (RLS)) comme le filtrage de la variation du filtre optimal et le bruit d'estimation (induit par le bruit de mesure). L'opération de filtrage effectuée par l'algorithme RLS dépend de la fenêtre utilisée dans le critère des moindres carrés. Pour pouvoir formuler un algorithme moindre carré récursif, il faut que la fenêtre puisse être exprimée d'une manière récursive. En pratique, seulement deux fenêtres ont été examiné (chartérisée chacune par un unique paramètre): la fenêtre exponentielle (W-RLS) et la fenêtre rectangulaire (SWC-RLS). Cependant, la fenêtre rectangulaire peut être généralisée, à un coût réduit, à une fenêtre généralisée (GSW-RLS) avec trois paramètres au lieu d'un seul, incluant les deux précédentes fenêtres comme cas spéciaux. Puisque la complexité de la fenêtre rectangulaire (SWC-RLS) est essentiellement le double de la fenêtre exponentielle (W-RLS), il est généralement admit que cette augmentation de la complexité entrane une amélioration de la poursuite. Nous prouvons que, avec un bruit d'estimation égal, les performances de la fenêtre exponentielle W-RLS surpasse généralement les performances de la fenêtre rectangulaire (SWC-RLS) dans le cas d'un problème de poursuite causal. Les performances de la fenêtre généralisée GSW-RLS sont généralement meilleures. Pour le problème de poursuite non-causal, le SWC-RLS est de loin le meilleur (la GSW-RLS tends vers la SWC-RLS). En présence d'un a priori statistique sur le canal, les paramètres des fenêtres sont estimés en minimisant l'erreur quadratique moyenne (MSE). Si on suppose que le spectre de variation du canal est un spectre plat, passe-bas; le GSW-RLS surpasse le W-RLS (qui surpasse son tour le SWC-RLS). Dans le cas générale, nous dérivons les expressions des fenêtres optimales pour la poursuite causale et non-causale. Il en ressort que la fenêtre exponentielle est optimale pour une poursuite causale d'un canal AR(1) ; tandis que la fenêtre rectangulaire est optimale pour la poursuite non-causale d'un processus de saut blanc.

Part I

Praclical Approaches on Bayesian Adaptive Filtering

Introduction and Motivations

Adaptive filtering have been extensively studied for a large range of applications including channel estimation, adaptive equalization, echocancelation, etc in a variety of stationary environment. For the nonstationary environments, two different classes of input have been studied for adaptive filtering algorithms. It has been shown that the Wiener solution has a time-varying characteristic. In contrast to adaptive filter convergence, which is a transient phenomenon, the tracking characteristics of the adaptive filter are known be to a steady state property of the filter. Consequently, good convergence properties do not ensure good tracking performance, and a compromise between the two properties are required for applications in a non-stationary environment. The standard adaptive filtering (SAF) such as the least mean-square (LMS) algorithm , and the recursive least-squares (RLS) algorithm are established as the principal algorithms to track for linear adaptive filtering. The convergence behaviors of both of these algorithms can be found in the literature ([1]), ([22]), ([10]). The RLS algorithm has a faster rate of convergence than the LMS algorithm and is not sensitive to variations in the eigenvalues of correlation matrix of the input signal. However, when operating in a non-stationary environment, they pocess only one parameter to adjust the tracking. Most of the work on adapting tractive capability has focused on adapting one tracking parameter. In RLS, it doe cost any computational complexity to make the forgetting factor time-varying. Modifications to fast RLS algorithms to allow a time-varying forgetting factor, as well as algorithms to adjust this forgetting factor on the basis of correlation matching have been pursued in [82]. The equivalent development for LMS algorithms concerns Variable StepSize (VSS) algorithms. Important developments were presented in [37],[39], [65],[64],[66] [38],[42]. Most of the VSS algorithms use the steepest-descent strategy and the instantaneous squared error cost function of the LMS

algorithm to adjust the additional parameter, which is the step-size. A related but different approach consists in running various adaptive filters with different time constants and selecting or combining their outputs, similarly to what is done in model order selection, see [71],[70], [68],[69].

A further refinement is to allow different tracking bandwidths for different filter components as is done in [40] with a VSS per filter coefficient and in [81] where the tracking capacity increases with frequency for the various frequency domain components of the filter. The work in [40] essentially shows that a "diagonal" state-space model may allow a simplification of the Kalman Filter (KF) to a LMS algorithm with a VSS per tap, but no attempt is made to automatically adjust the resulting stepsizes.
Besides the statistical modeling of the parameter variation, another important ingredient in Bayesian adaptive filtering is the incorporation of prior knowledge on the coefficient sizes.
The influence of the prior distribution on Bayesian estimation, depends on the confidence on the observation, which in turn depends on the length of the observation, and on the SNR. In general, as the number of the observation samples and the SNR increase, the variance of the estimate, and the influence of the prior, decrease. In estimating a Gaussian distributed parameter observed in AWGN, as the length of the observation N increases, the importance of the prior decrease, and the MAP estimate tends to the ML estimate.
Indeed, when tracking time-varying filters, it becomes possible to learn the variances of the filter coefficients. This aspect has been exploited for a while in a rudimentary, binary form for sparse filters: filter coefficients are either adapted or deemed to small and kept zero (for each filter coefficient, the stepsize is either 0 or a constant). More recently, a smoother evolution of the stepsize has been introduced, leading to the Proportionate LMS (PLMS) algorithm, motivated e.g. by acoustic echo cancellation in which the adaptive filter has many coefficients, but their value tapers off, see [78],[79]. Similar prior information is starting to be taken into account for (LMMSE) channel estimation in wireless communications [84], where the evolution of the channel coefficient variances along the impulse response is called the power delay profile.

The time variation of the optimal filter can be described by either expanding the filter coefficients into fixed time-varying (e.g. sinusoidal) basis functions (basis expansion

models (BEMs)) [24], or by modeling [10], them as stationary processes. The latter approach is perhaps better suited for minimum delay online processing. This case of constant slow variation of the filter coefficients (”drifting” parameters) is to be contrasted with another possible case of only occasional but significant variation (”jumping” parameters) which shall not be considered here. A lot of work has been done on optimizing the single parameter regulating the tracking speed of classical LMS or exponentially weighted RLS algorithms [1],[6],. For LMS, such an adaptive optimization leads to the class of Variable Step-Size (VSS) algorithms, see e.g.[40] and references therein. Adaptive filtering algorithms with a single adaptation parameter do not take into account that different portions of the filter may have different variation speeds and/or different magnitudes and hence can be quite suboptimal. One noteworthy attempt to overcome this limitation is the introduction of a coefficient-wise VSS, but the automatic adaptation of these VSSs is a difficult task. In Bayesian Adaptive Filtering (BAF)[12], prior information on the filter coefficient variances and variation spectra is exploited to optimize adaptive filter performance. A straightforward way to implement BAF is to use the Kalman filter. However, the complexity of the Kalman filter is much higher compared to that of the popular LMS adaptive filtering algorithm. Furthermore, the Kalman filter needs to be augmented with a state-space model identification technique.

In order to design a general Bayesian Adaptive Filtering (BAF) model, one can proceed in two basic approaches.

It has been known for a long time that for best tracking results adaptive filtering should be formulated as a Kalman filtering problem, leading to Bayesian Adaptive Filtering (BAF). BAF techniques with acceptable complexity can be obtained by focusing on a diagonal AR(1) model for the time-varying optimal filter settings. The hyper-parameters of the AR(1) model can be adapted by introducing EM techniques and one sample fixed-lag smoothing at little extra cost. Standard AF techniques such as the LMS and RLS algorithms are equipped with only one hyper-parameter (stepsize, forgetting factor) to optimize their tracking behavior.

Thesis outline and contributions

While adaptive filtering is in principle intended for tracking non-stationary systems, most adaptive filtering algorithms have been designed for converging to a fixed unknown filter. When actually confronted with a non-stationary environment, they possess just only one parameter (stepsize, forgetting factor) to adjust their tracking capability. Virtually the only existing optimal approach is the Kalman filter, in which the time-varying optimal filter is modeled as a vector AR(1) process. The Kalman filter is in practice never applied as an adaptive filter because of its complexity and large number of unknown parameters in its state-space (AR(1)) model. This motivated our work in this thesis to look for practical techniques to take advantage of the Kalman optimality with reduced complexity to make this approach applicable. Also to propose different methods for the estimation of the large number of parameters.
To better understand the different parameters used in the first part of the thesis we conside we begin in chapter 1 by a brief introductory presentation of state space model and general principals of standard adaptive filtering. We then introduce critical parameters defining the wireless channel medium (Power-Delay Profile, Doppler Bandwidth,...) followed by an illustrative example.

Chapter 2 focuses on Kalman smoothers and Kalman filters which form an important component of algorithms developed later in this thesis. While most of the results presented in this chapter are well-known to much of the adaptive filtering research, some results, such as the recursion equation and the fixed interval Kalman smoother, are either new or obscure in origin. Even in the case of the better-known material, its importance to this thesis merits its restatement. Following an introductory section on notation, we present a derivation of a recursive Kalman filter and include a practical strategy for ensuring that the results of the filter are numerically stable. An identical approach is then taken for the Kalman smoother. Our attention is then directed toward the stability of the time-varying Kalman filter and smoother.

Chapter 3 is concerned primarily with the problem of selecting a model from a set of

candidates. In the ensuing text we shall assume that the experiment design, data collection and model structure selection operations have already been performed. In addition, we shall assume that the model structure is a parametric one (that is the model structure is parameterized by a finite-dimensional vector of real numbers). The task of model selection, therefore, is equivalent to that of selecting a suitable parameter vector from a set of candidates. This practice is known as " Parameter Estimation" .

In chapter 4, we introduce a means of maximum-likelihood estimation of parameters that is applicable in many cases when direct access to the necessary to make the estimates is impossible , or when some of the data are missing. Such inaccessible data are present, for example, when an outcome is a result of an accumulation of simpler outcomes, or when outcomes are clumped together (e.g., in a inning or histogram operation). There may also be data dropouts or clustering such that the number of underlying data points in unknown (censoring and/or truncation). The EM (expectation-maximization) algorithm is ideally suited to problems of this sort, in that it produces maximum-likelihood (ML) estimates of parameters when there is a many-to-one mapping from an underlying distribution to the distribution governing the observation. The EM algorithm consist of two primary steps: an expectation step, followed by maximization step. The expectation is obtained with respect to the unknown underlying variables, using the current estimate of the parameters and conditioned upon the observations. The maximization step then provides a new estimate of the parameters. These two step are iterated until convergence.

We tackle in chapter 5 the problem of filtering in non-stationary environments. We first begin by an introductory state of the art review which includes the existent algorithms on adaptive filtering (LMS and LRS). These algorithms experience performances limitation in terms of tracking and convergence in non-stationary environments. This motivates our work to propose more efficient techniques for such Bayesian techniques. Our proposed methods take into consideration a priori information about the system variations such as the PDP, Doppler bandwidth, ...etc. We thus propose two different approaches, the first one is based on Wiener Filtering (WF) while the other one is based on Kalman Filtering (KF). In this chapter we develop the first approach and the second one will be

considered in the next chapter. The proposed algorithms can be applied in many situations and as an example we consider in this chapter its application for system identifications in particular mobile radio channel for the importance of this medium in wireless communications. We provide numerical results that show the proposed algorithm advantage compared to existent algorithms in terms of Excess Mean Squared Error (EMSE) for different PDPs and Doppler shifts.

In chapter 6, we continue our study of the Bayesian Adaptive Filtering (BAF) concept that we introduced in the previous chapter. The proposed technique is based on modeling the optimal adaptive filter coefficients as a stationary vector process, in particular as a AR(1) model. Optimal adaptive filtering with such a state model becomes Kalman filtering. The complexity of the resulting algorithm is $O(N^3)$ and in order to reduce this complexity we propose a diagonal AR(1) based model, of complexity $O(N^2)$ which is comparable to RLS complexity. For the AR(1) model parameters estimation, we propose an adaptive version of the EM algorithm with complexity limited to $O(N)$ for the EM part. The proposed parameters estimation method leads to linear prediction on reconstructed optimal filter correlations, and hence a meaningful approximation/estimation compromise. To further reduce the initial adaptive EM-Kalman algorithm complexity, we develop a second approach based on component-wise EM-Kalman (This technique is of complexity $O(N)$ which is comparable to LMS complexity). To compare the proposed algorithms performance, we derived the exact analytical expressions of Excess Mean Square Error (EMSE) in the steady-state in the general case and we proposed a comparison for the application to radio mobile communications where the priori information is the fading, PDP and the Doppler shift. The former proposed algorithm is outperformed by the EM-Kalman based approach, in terms of tracking and convergence. To offer comparable performance with the AR(1) algorithm with same complexity of component-wise EM, we propose in the following chapter a two-stage approach.
Up to this point, we propose different Bayesian techniques with different complexities. Thus we proposed a EM-Kalman algorithm with complexity $O(N^2)$. To reduce the complexity, we presented, the adaptive component-wise EM-Kalman algorithm with complexity $O(N)$ but which shows performance limitations in terms of tracking and convergence compared the previous technique. This motivated our study for the development

of another technique with the same performance as the EM-Kalman but with the same complexity as the component-wise EM-Kalman in chapter 7. The proposed two-stage approach consists of a first step employing a basic fast tracking adaptive filter, followed by lowpass filtering and downsampling of the time-varying filter coefficients. The second step then applies Kalman filtering at the reduced rate on a simplified state-space model, with an additive white noise measurement equation. The parameters in the state equation can be conveniently identified with an adaptive EM algorithm. The first stage would typically employ a (Normalized) LMS algorithm with a large stepsize. The main assumption underlying the proposed two-stage approach is that even in fast tracking applications, the bandwidth of the optimal filter variation is typically small compared to the signal bandwidth, motivating the downsampling operation. The first stage attempts to provide a bias-free filter estimate whereas the second stage optimizes the estimation variance.

In the second part of the thesis we consider the problem of window optimization issues in recursive Least-Squares adaptive filtering and tracking. We consider tracking of an optimal filter modeled as a stationary vector process. We interpret the Recursive Least-Squares (RLS) adaptive filtering algorithm as a filtering operation on the optimal filter process and the intantaneous gradient noise (induced by the measurement noise). The filtering operation carried out by the RLS algorithm depends on the window used in the least-squares criterion. To arrive at a recursive LS algorithm requires that the window impulse response can be expressed recursively (output of an IIR filter). In practice, only two popular window choices exist (with each one tuning parameter): the exponential weighting (W-RLS) and the rectangular window (SWC-RLS). However, the rectangular window can be generalized at a small cost for the resulting RLS algorithm to a window with three parameters (GSW-RLS) instead of just one, encompassing both SWC- and W-RLS as special cases. Since the complexity of SWC-RLS essentially doubles with respect to W-RLS, it is generally believed that this increase in complexity allows for some improvement in tracking performance. We show that, with equal estimation noise, W-RLS generally outperforms SWC-RLS in causal tracking, with GSW-RLS still performing better, whereas for non-causal tracking SWC-RLS is by far the best (with GSW-RLS not being able to improve). When the window parameters are optimized for causal tracking MSE, GSW-RLS outperforms W-RLS which outperforms SWC-RLS. We also derive the optimal window

shapes for causal and non-causal tracking of arbitrary variation spectra. It turns outs that W-RLS is optimal for causal tracking of AR(1) parameter variations whereas SWC-RLS if optimal for non-causal tracking of integrated white jumping parameters, all optimal filter parameters having proportional variation spectra in both cases.
We should mention that the notation of the two parts are different.
We present our conclusions and perspectives in chapter 9.

The results of this thesis have been published in the following papers:

- T. Sadiki, D. T. M. Slock, "Steady-state performance analysis of Bayesian adaptive filtering," ISCCSP 2006, IEEE International Symposium on Communication, Control and Signal Processing. 13-15 March, 2006, Marrakech, Morroco.

- T. Sadiki, D. T. M. Slock "Bayesian adaptive filtering: principles and practical approaches," EUSIPCO 2004, 12th European Signal Processing Conference, September 6-10, 2004, Vienna, Austria.

- T. Sadiki, D. T. M. Slock, "Steady-state comparison between Bayesian adaptive filtering and standard adaptive filtering" Asilomar 2006, 40th IEEE Annual Asilomar Conference on Signals, Systems and Computers, November 7-10, 2004, Pacific Grove, USA.

- T. Sadiki, D. T. M. Slock, " performance analysis of two-stage approach to Bayesian adaptive filtering," Submitted to ICASSP 2007, USA.

- T. Sadiki, D. T. M. Slock, "Bayesian adaptive filtering at linear cost" SSP 2005, 13 th IEEE Workshop on Statistical Signal Processing, July 17-20, 2005, Bordeaux, France.

- T. Sadiki, D. T. M. Slock, "On the convergence of bayesian adaptive filtering" ISSPA 2005, 8th International Symposium on Signal Processing and Its Applications, August 29-September 1, 2005, Sydney, Australia.

- T. Sadiki, D. T. M. Slock, "Low complexity bayesian adaptive filtering with independent AR(1) filter coefficient models " SSP 2005, 13 th IEEE Workshop on Statistical Signal Processing, July 17-20, 2005, Bordeaux, France.

- T. Sadiki, D. T. M. Slock, "Bayesian adaptive filtering at linear cost" SSP 2005, 13 th IEEE Workshop on Statistical Signal Processing, July 17-20, 2005, Bordeaux, France.

- T. Sadiki, D. T. M. Slock, "A two-stage approach to Bayesian adaptive filtering" ISCCSP 2006, IEEE International Symposium on Communication, Control and Signal Processing. 13-15 March, 2006, Marrakech, Morroco.

- T. Sadiki, M. Triki, D. T. M. Slock, "Window optimization issues in recursive least-squares adaptive filtering and tracking" Asilomar 2004, 38th IEEE Annual Asilomar Conference on Signals, Systems and Computers, November 7-10, 2004, Pacific Grove, USA.

- T. Sadiki, M. Triki, D. T. M. Slock, "Window optimization issues in recursive least-squares adaptive filtering and tracking," Submitted to IEEE transactions on signal processing.

- T. Sadiki, D. T. M. Slock, "Practical Approaches on Bayesian Adaptive Filtering" Submitted to IEEE transactions on signal processing.

Chapter 1

Signals and Systems Preliminaries

The purpose of this chapter is to define a set of notation used in later chapters and to present a theoretical background to some of the problems tackled in those chapters.

1.1 Linear State-Space Systems

State-space equations provide a compact, flexible and attractive method of modeling system behavior. They represent, in a sense, a highly satisfactory modeling approach due to the fact that they allow a complete description of the internal as well as external characteristics of a given system. These positive attributes have encouraged, at least in the case of Linear time varying (LTV) systems, the emergence of a rich theoretical underpinning for further developments in this thesis. An $n-th$ order, linear, time-varying, discrete-time system may be described by the following stochastic state-space model.
Let consider the system driven by noise, with noisy observation

$$\begin{aligned} H_k &= AH_{k-1} + W_k \\ y_k &= X_k^H H_{k-1} + v_k \end{aligned} \tag{1.1}$$

As indicated, the state-variable system is time-varying. The following assumption are made about system:

- The sate noise process W_k is zero-mean, with covariance

$$E[W_k W_k^H] = Q \tag{1.2}$$

  the noise is uncorrelated among samples. For the Bayesian approach, we assume that W_k is Gaussian.

- The observation noise v_k is zero-mean, with covariance

$$E[v_k v_k^H] = \sigma_v^2 \tag{1.3}$$

  the noise is uncorrelated among samples. For the Bayesian approach, we assume that v_k is Gaussian.

- where $H_k \in R^N$ is the state vector.

- $y_k \in R^m$ is the output signal and , $X \in R^N$ the input.

Here we have used the convention that a subscript k denotes the value of a signal at the $k-th$ sampling instant.

For the system outlined above we can ensure that the state and output sequences always exist for bounded inputs (that is, the outputs and states converge) by requiring that the system be strictly stable . This is described by Property (1).

Property 1 *(Strict Stability). A discrete linear time-varying system*

$$H_k = AH_{k-1} + W_k \tag{1.4}$$

is strictly stable if and only if the eigenvalues of $A \in R^{N \times N}$ are in the open unit disc. That is, $\mid \lambda_i(A) \mid$, $\forall i = 1, 2, \ldots, N$.

Controllability and observability are two system properties which will be important for the development of the ensuing theory. These concepts are widely known and discussions of them are commonly available in control literature. For further information see, for example, Goodwin and Sin [43]. The following definitions address the properties of systems modeled by the state-space equations (Strict Stability). A discrete linear time-varying system

$$\begin{aligned} H_k &= AH_{k-1} + W_k \\ y_k &= X_k^H H_{k-1} + v_k \end{aligned} \tag{1.5}$$

1.2 Deterministic and Stochastic Signals

It is often convenient for analytical reasons to discuss the spectral properties of various time-domain signals. However, not all signals may be said to possess a spectrum. In this section we discuss two types of processes - stationary stochastic processes and a generalization of them, known as quasi-stationary processes. The latter allows for the possibility of deterministic components. Both of these types of processes have spectra. First, though, we need to introduce the concept of expected values.

Definition 2 ***(Expected value)****. When it exists, the expectation (or expected value) of a continuous random variable x, denoted $E\{x\}$, is defined by*

$$E\{x\} = \int_{-\infty}^{+\infty} xp(x)dx, \tag{1.6}$$

where $p(x)$ is the probability density function of x. $E\{x\}$ is said to exist if $E\{\mid x \mid\} < 1$.

Definition 3 ***(Conditional expected value)****. When it exists, the conditional expected value of the random variable x given the random variable y is defined as $E\{x \mid y\}$*

$$E\{x \mid y\} = \int_{-\infty}^{+\infty} xp(x \mid y)dx, \tag{1.7}$$

where $p_y(x) \triangleq p(x \mid y)$ is the conditional probability density function of x given y. The latter is defined by

$$p(x \mid y) = \frac{p(x,y)}{p(y)}. \tag{1.8}$$

Remark 4 *Often, in the interests of notational convenience we shall denote conditional dependence by subscripting. For example, equation (1.9) may be written as*

$$E_y\{x\} = \int_{-\infty}^{+\infty} x p_y(x) dx, \tag{1.9}$$

What follows is a brief treatment of stationary processes. There is a great deal of existing theory for systems (particularly linear ones) of this type. For further information see, for example, [146], or for a more introductory, Anderson and Moore [153] or Papoulis [152].

1.2.1 Stationary Stochastic Process

A random process $\{z_k\}$ is said to be strict-sense stationary (SSS) if its statistical properties, such as the joint distribution $p(z_k, z_{k+1}, \ldots, z_{k+n})$ for any n, are invariant to a shift of the origin. It is wide-sense stationary (WSS) provided that

$$E[z_k] = \mu, \forall t$$

$$E[z_{k+\tau} z_k^T] = R_z(\tau), \forall t \ \ \forall \tau$$

where $E\{\}$ is the expectation operator defined by Definition (2) It arises (see [152] that a Gaussian (that is, Normally-distributed) WSS process is also SSS since the mean and auto-correlation properties are sufficient to specify the distribution uniquely.

1.3 Standard Adaptive Filtering

1.3.1 Least Mean Square (LMS) Algorithm

The least mean square (LMS) is a search algorithm in which a simplification of the gradient vector computation is made possible by appropriately modifying the objective function [1],[3]. The LMS algorithm, as well as others related to it, is widely used in various applications of adaptive filtering due to its computational simplicity [4]. The convergence

characteristics of the LMS are examined in order to establish a range for the convergence factor that will guarantee stability. The convergence speed of the LMS is shown to be dependent of the eigenvalue spread of the input-signal correlation matrix [1],[3]. The LMS algorithm is by far the most widely used algorithm in adaptive filtering for several reasons. The main features that attracted the use the LMS algorithm are low computational complexity, proof of convergence in stationary environment, unbiased convergence in the mean to the Wiener solution, and stable behavior when implemented with finite-precession arithmetic.

The optimal solution for the parameters of the adaptive filter leads to the minimum mean-square error (MMSE) in estimating the reference signal y_k in Eq.(1.5). The optimal Wiener solution [1] is given by

$$H = R^{-1}p \tag{1.10}$$

where $R = E[X_k X_k^H]$ and $p = E[d_k X_k]$, assuming that y_k and X_k are jointly wide-sense stationary.

If good estimates of matrix R denotes by $\hat{R}_k$, and of vector p, denoted by $\hat{p}_k$, are available, a steepset-descent-based algorithm can be used to search the Wiener solution equation (1.10) as fellows:

$$\begin{aligned} \hat{H}_k &= \hat{H}_{k-1} - \mu \hat{g}_{k,\hat{H}} \\ &= \hat{H}_{k-1} + 2\mu(\hat{p}_k - \hat{R}_k \hat{H}_{k-1}) \end{aligned} \tag{1.11}$$

for $k = 0, 1, \ldots,$ where $\hat{g}_{k,\hat{H}}$ represents an estimate of the gradient vector of the objective function with respect to the filter coefficients.

One possible solution is to estimate the gradient vector by employing instantaneous estimates for R and p as fellows:

$$\begin{aligned} \hat{R}_k &= X_k X_k^H \\ \hat{p}_k &= y_k X_k \end{aligned} \tag{1.12}$$

The resulting gradient estimate is given by

$$\begin{aligned} \hat{g}_{k,\hat{H}} &= -2y_k X_k + 2X_k X_k^H \hat{H}_{k-1} \\ &= 2X_k(-y_k + X_k^H \hat{H}_{k-1}) \\ &= -2e_k X_k \end{aligned} \tag{1.13}$$

Note that if the objective function is replaced by the instantaneous square error e_k^2, instead of the MSE, the gradient estimate above represents the true gradient vector since

$$\begin{aligned}\frac{\partial e_k^2}{\partial \hat{H}} &= \left[2e_k\frac{\partial e_k}{\partial H_0}\, 2e_k\frac{\partial e_k}{\partial H_1}\, \ldots 2e_k\frac{\partial e_k}{\partial H_N}\,\right]^T \\ &= -2e_k X_k \\ &= \hat{g}_{k,\hat{H}} \end{aligned} \tag{1.14}$$

The resulting gradient-based algorithm is known, because it minimizes the mean of the squared error, as the least-mean-square (LMS) algorithm, whose updating equation is

$$\hat{H}_k = \hat{H}_{k-1} + 2\mu e_k X_k \tag{1.15}$$

where the convergence factor μ should be chosen in range to guarantee convergence.

1.3.2 Recursive Least-Square (RLS) Algorithm

Least-squares algorithms aim at the minimization of the sum of the squares of the difference between the desired signal and the model filter output [1],[3]. When new samples of the incoming signals are received at every iteration, the solution for the least-squares problem can be computed in recursive form resulting in the recursive least-squares (RLS) algorithms.

The RLS algorithms are known to pursue fast convergence even when the eigenvalue spread of the input signal correlation matrix is large. These algorithms have excellent performance when working in time-varying environments. All these advantages come with the cost of an increased computational complexity and stability problems, which are not as critical in LMS algorithms.

The objective her is to choose the coefficients of the adaptive filter such that the output y_k, will match the desired signal as closely as possible in the least-squares sens. The minimization process requires the information of the input signal available so far. Also, the objective function we seek to minimize is deterministic.

The generic FIR adaptive filter realized in the direct form is shown in Fig.1.1.

The input signal information vector at a given instant k is given by

$$X_k = [x_k\, x_{k-1} \ldots\, x_{k-N}]^T$$

wher N is the order of the filter. The coefficients $\hat{h_{jk}}$, for $j = 0, 1, \ldots, N$, are adapted aiming at the minimization of a given objective function. In the case of least-squares algorithms, the objective function is deterministic and is given by

$$\begin{aligned} \epsilon_k &= \sum_{i=1}^{k} \lambda^{k-i} e_i^2 \\ &= \sum_{i=1}^{k} \lambda^{k-i} [y_i - X_i^H \hat{H}_{k-1}]^2 \end{aligned} \tag{1.16}$$

where e_i is the output error at instant $i - th$ and $\hat{H}_{k-1}$ is the adaptive filter coefficient vector. The parameter λ is an exponential weighting factor that should be chosen in the range $0 \leq \lambda < 1$. This parameter is also called forgetting factor since the information of the distant past has an increasingly negligible effect on the coefficient updating.
As can be noted, each error consists of the difference between the desired signal and the filter output, using the most recent coefficients $\hat{H}_{k-1}$. By differentiating ϵ_k with respect to $\hat{H}_{k-1}$, it follows that

$$\frac{\partial \epsilon_k}{\partial \hat{H}_{k-1}} = -2 \sum_{i=1}^{k} \lambda^{k-i} [y_i - X_i^H \hat{H}_{k-1}] \tag{1.17}$$

By equating the result to zero, it is possible to find the optimal vector $\hat{H}_{k-1}$ that minimizes the least-squares error, through the following relation:

$$-\sum_{\mathbf{i=1}}^{\mathbf{k}} \lambda^{\mathbf{k-i}} \mathbf{X_i} \mathbf{X_i^H} \hat{\mathbf{H}}_{\mathbf{k-1}} + \sum_{\mathbf{i=1}}^{\mathbf{k}} \lambda^{\mathbf{k-i}} \mathbf{X_i} \mathbf{y_i} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The resulting expression for the optimal coefficient $\hat{H}_k$ is given by

$$\begin{aligned} \hat{H}_k &= \big[\sum_{i=1}^{k} \lambda^{k-i} X_i X_i^H\big]^{-1} \sum_{i=1}^{k} \lambda^{k-i} X_i y_i \\ &= R_D^1(k) P^D(k) \end{aligned} \tag{1.18}$$

where $R_D^1(k)$ and $P^D(k)$ are called the deterministic correlation matrix of the input signal and deterministic cross-correlation vector between the input and desired signals, respectively.

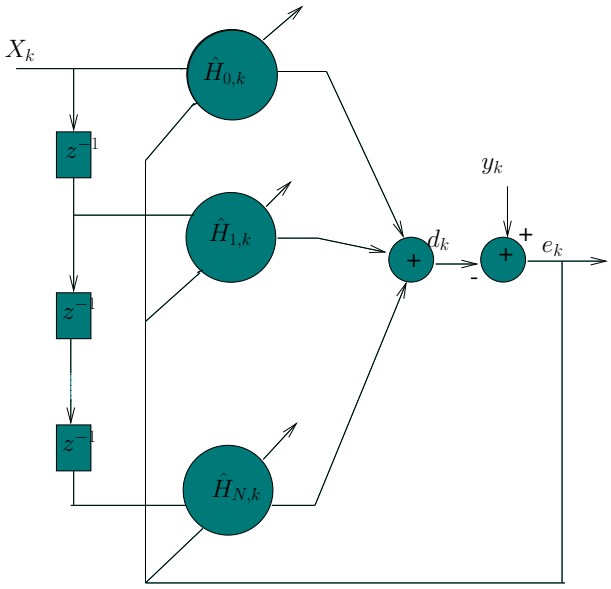

Figure 1.1: Adaptive FIR filter

In equation (1.18) it was assumed that $R_D(k)$ is nonsingular. However if $R_D(k)$ is singular a generalized inverse [1] should be used instead in order to obtain a solution for $\hat{H}_k$ that minimizes ϵ_k. Since we are assuming that in most practical applications the input signal has persistence of excitation, the cases requiring generalized inverse are not discussed here. It should be mentioned that if the input signal is considered to be zero for $k < 0$ then $R_D(k)$ will always be singular for $k < N$, i.e., during the initialization period. During this period, the optimal value of the coefficients can be calculated for example by the back substitution algorithm [book].

The straightforward computation of the inverse of $R_D(k)$ results in an algorithm with computational complexity $O(N^3)$. In the conventional RLS algorithm the computation of the inverse matrix is avoided through the use of the matrix inversion lemma [1]. Using the matrix inverse lemma, the inverse of the deterministic correlation matrix can then be calculated in the following from

$$S_k = R_D^{-1}(k) = \frac{1}{\lambda}\left[S_{k-1} - \frac{S_{k-1}X_kX_k^HS_{k-1}}{\lambda + X_k^HS_{k-1}X_k}\right] \tag{1.19}$$

The complete conventional RLS algorithm is described below.

Conventional RLS algorithm
- Initialization
$S_D(-1) = \delta I$
where δ can be the inverse of the input-siganl power estimate
$S_D(-1) = X_{-1} = [0\,0\,\ldots\,0]^T$
Do for $k \geq 0$:
$S_k = R_D^{-1}(k) = \frac{1}{\lambda}\left[S_{k-1} - \frac{S_{k-1}X_k X_k^H S_{k-1}}{\lambda + X_k^H S_{k-1} X_k}\right]$
$P_D(k) = \lambda P_D(k-1) + y_k X_k$
$\hat{H}_k = S_k P_D(k)$
If necessary compute:
$d_k = \hat{H}_k^H X^k$
$e_k = y_k - d_k$

1.4 Application

The type of application is defined by the choice of the signal acquired from the environment to be the input and desired-output signals. The number of different applications in which adaptive techniques are being successfully used has increased enormously during the last decade. Some examples are echo cancellation, equalization of dispersive channel, signal enhancement, noise cancelling and system identification. The study of different applications is not the main scope of the thesis. However, some applications are considered, like system identification, in particular **Mobile Radio Channel (MRC)**.

1.5 Example: Multi-path Channel Model

The physical basis of multi-path propagation is given by the reception of multiple copies of the transmitted signal, each having traveled along a different propagation path. In a typical environment, each propagation path has a different length and, thus, the signal copy having traveled along this path arrives at the receiver with a different delay. Signal

copies traveling along short paths will arrive earlier, while other copies traveling along longer paths will arrive later. The channel is said to have a memory since it stores the signal copies for a certain time period, i.e., the duration of the propagation. Beside the different delays, the signal copies are attenuated differently, since along their different propagation paths they traverse different obstacles of different shapes and sizes. Moreover, the signal copies arrive at the receiver from different directions and with different phases. The superposition of all these differently delayed, attenuated, and phase-shifted signal copies at the receiver results in an interference pattern, which alternately behaves constructively and destructively. If nothing moves within the propagation environment, the received signal will remain constant, and therefore the channel is said to be time invariant. In contrast, if any kind of change is encountered in the propagation environment, all or some paths will change in time and, thus, the interference pattern will change in time. As a consequence the channel becomes time variant. The multi-path channel model is a mathematical model that is meant to account for all the effects of multi-path propagation. Let us first consider the transmission of a bandpass signal $s(t)$ at carrier frequency f_c in the case of a time invariant channel. By associating to each path p a different length l_p and a different attenuation A_p, the received bandpass signal $r(t)$ being the superposition of all copies can be written as

$$r(t) = \sum_p A_p s(t - \frac{l_p}{c}) \tag{1.20}$$

Considering the complex envelope representations, $s_e(t)$ and $r_e(t)$ of the bandpass signals $s(t)$ and $r(t)$ respectively, the input-output relationship given in Eq. (8.4) becomes

$$r_e(t) = \sum_p A_p e^{-j\phi_p} s_e(t - \tau_p) \tag{1.21}$$

where $\phi_p = \frac{2\pi f_c l_p}{c}$ and $\tau_p = \frac{l_p}{c}$ denote respectively the phase shift of the carrier frequency and the delay caused by the different length of path p, and c is the speed of light.

Thus, in a static environment, a multi-path propagation leads to the interference of multiple copies with different attenuations $\{A_p\}$, different phase shifts $\{\phi_p\}$, and different

delays $\{\tau_p\}$. The time invariant channel model can then be modeled as a *linear time invariant causal filter* with baseband impulse response

$$h_e(t) = \sum_p A_p e^{-j\phi_p} \delta(\tau - \tau_p) \tag{1.22}$$

where $\delta(.)$ denotes Dirac's delta function. Now let us consider the effect of motion in the channel. Let α_p denote the angle of arrival of path p with respect to the direction of motion of the receiver, as shown in Figure **??**

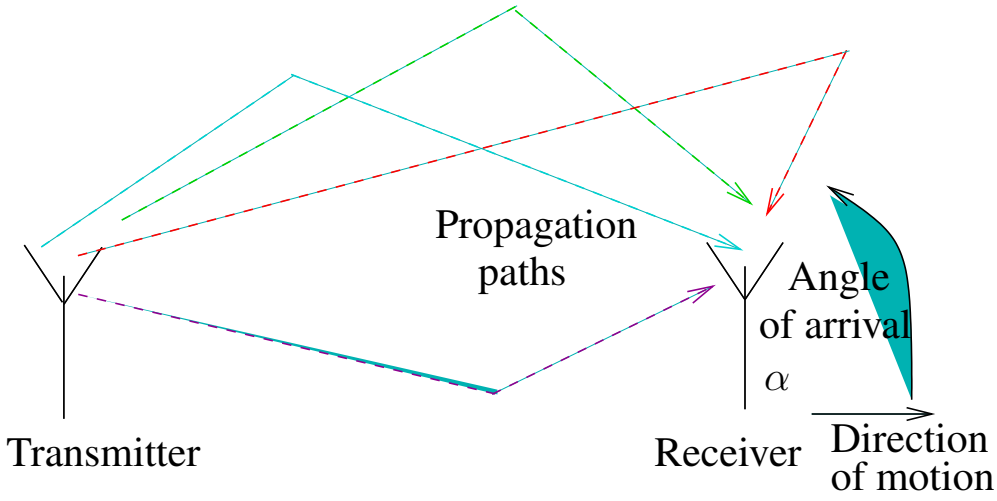

Figure 1.2: Design of signal arrival in a multi-path environment.

The path length is now variant and relates to mobile speed ν as [2]

$$l_p(t) = l_p(0) - \nu cos(\alpha_p)t \tag{1.23}$$

From (3.5), we obtain a different function for the complex envelope of the received signal, which now depends on time t, as given below

$$r_e(t) = \sum_p A_p e^{-j\phi_p} e^{j2\pi \frac{\nu}{c} cos(\alpha_p)t} S_e(t - \tau_p + \frac{\nu}{c} cos(\alpha_p)t) \tag{1.24}$$

Equation (3.6) can be simplified by making the three following operations. First, we regroup into the complex A_p the attenuation α_p and the term ϕ_p. Secondly the extra delay

caused by $l_p(t)$ compared to the delay τ_p caused by the path length $l_p(0)$. At last, we introduce the *Doppler frequency* $f_d = \frac{f_c \nu}{c}$ and the *Doppler shift* $\nu_p = cos(\alpha_p) f_d$. With, this we obtain the simplified form

$$r_e(t) = \sum_p A_p e^{j2\pi\nu_p t} S_e(\tau - \tau_p) \tag{1.25}$$

Thus, from (4.16), we can observe that motion introduces a frequency offset $\{\nu_p\}$ of the carrier in addition to the signal changes that are already present in static conditions. As for the time invariant channel, the time variant channel is modeled by a *linear time variant causal filter* with impulse response

$$h_e(t, \tau) = \sum_p A_p e^{j2\pi\nu_p t} \delta(\tau - \tau_p) \tag{1.26}$$

The received signal can therefore be expressed as the convolution of the transmitted signal with the impulse response $h_e(t, \tau)$ with respect to the time delay τ:

$$r_e(t) = \int_0^\infty h_e(t, \tau) s_e(t - \tau) d\tau \tag{1.27}$$

Equivalently to the time domain response $h_e(t, \tau)$, the channel can also be characterized in the frequency domain by the *time variant transfer function* $H(t, f)$, which is the Fourier transform of $h_e(t, \tau)$ with respect to τ. It is obtained from

$$H(t, f) = \sum_p A_p e^{j2\pi(\nu_p t - f\tau_p} \tag{1.28}$$

In the frequency domain, the spectra of transmitted and received signals are related by simple multiplication with the transfer function $H(t, f)$:

$$R(t, f) = H(t, f) S(f) \tag{1.29}$$

Thus, the transfer function $H(t, f)$ determines the attenuation experienced at time t by the spectrum component $S(f)$ at frequency f.

For a large number of paths in the propagation environment and in the absence of a LOS component, the central limit theorem applies to the time variant transfer function $H(t, f)$ and justifies its Gaussian distribution in both time and frequency. By using polar coordinates, the amplitude ρ of $H(t, f)$ thus follows a Rayleigh distribution while its phase θ follows a uniform distribution in $[0, 2\pi[$ [2]. In the presence of a LOS component, however, a Rice distribution is more generally assumed for the amplitude ρ of $H(t, f)$. On the other hand, for a small number of paths, the assumption of Gaussian random process as a result of the central limit theorem is The Mobile Radio Channel no more appropriate. In this case,ρ is generally assumed to follow a Nakagami distribution either in the presence or absence of a LOS component [2].

The Rayleigh distribution shows up in most non-LOS scenarios, which are encountered mostly in indoor and macro-cellular urban environments. In these scenarios, the performance of communication systems are worse than in scenarios where Rice distribution applies. This is because Rice fading is less destructive than Rayleigh fading. In this thesis, Rayleigh fading is assumed since our focus is on macro-cellular urban environments.

Characterization in Time and Frequency As shown in (3.11), the Doppler shifts $\{\nu_p\}$ and time delays $\{\tau_p\}$ are responsible of the time and frequency variations of the attenuation experienced by the received signal. Although Doppler shifts are frequency offsets of the carrier frequency that may induce *Inter-Carrier Interference (ICI)* in multi-carrier systems, their overall impact on the received signal is interpreted as a time selective behavior. For the time delays it is the opposite. While the delays are time offsets of the transmitted signal that may induce *Inter-Symbol Interference* (ISI), their impact on the received signal is interpreted as a frequency selective behavior. Two quantities are commonly used in practice to describe the impact of time delays and Doppler shifts on the received signal. They are the *delay spread* $\Delta\tau$ and *Doppler spread* $\Delta\nu$. The delay spread relates to the frequency selectivity of the channel, whereas the Doppler spread relates to the time selectivity of the channel.

Delay Spread and Frequency Selectivity The delay spread describes the time spread of the signal caused by multi-path propagation with several paths of different lengths and, thus, of different delays. Since the delays $\{\tau_p\}$ are different for different paths, the transfer function $H(t, f)$ will then vary with respect to frequency f and the spectrum $S(f)$ will undergo different attenuations for different frequency components. This phenomenon is referred to as *frequency selectivity*.

The severity of the frequency selectivity is measured by the product of the delay spread $\Delta\tau$ with the bandwidth W of the signal. So, if the delay spread is small compared to the inverse of W, that is the symbol time T_s, which corresponds to values of $W\Delta\tau$ smaller than 1, the transfer function $H(t, f)$ is nearly constant within the bandwidth W and all frequency components of the spectrum $S(f)$ will then have almost the same attenuation. In this case the channel is said to be flat or frequency non selective. On the other hand, if the delay spread is significant compared to symbol time, which corresponds to values of $W\Delta\tau$ greater than 1, the transfer function varies within the bandwidth W and the frequency components of the signal will be differently attenuated. Here, the channel is said to be frequency selective and the receiver suffers in the time domain from ISI. A detailed picture of the frequency selectivity of the multi-path channel is given by the spaced frequency correlation function of $H(t, f)$. This function gives us the correlation between the transfer function at different frequencies and it is given by [2]

$$\Phi_H(\Delta f) = \frac{1}{2}E[H(t, f)H^*(t, f - \Delta f)] \tag{1.30}$$

Substituting of (3.11) in (3.13 yields

$$\Phi_H(\Delta f) = \frac{1}{2}E[\sum_p \sum_q A_p A_q^* e^{j2\pi(\nu_p-\nu_q)t} e^{-j2\pi f(\tau_p-\tau_q)t} e^{-j2\pi\Delta f\tau_q}] \tag{1.31}$$

If scatterers at different delays $\{\tau_p\}$ are uncorrelated, the autocorrelation function depends only on the frequency spacing Δf. This assumption is called the *Uncorrelated Scattering (US)* assumption of multi-path channels. It can be written as

$$\frac{1}{2}E[A_pA_q^*e^{j2\pi(\nu_p-\nu_q)t}] = \sigma_p^2\delta_{pq} \tag{1.32}$$

The variance σ_p^2 is the average power of the $p-th$ signal copy. From (3.15), the spaced frequency correlation function in (3.14) turns into

$$\Phi_H(\Delta f) = \sum_p \sigma_p^2 e^{-j2\pi\Delta f\tau_p} \tag{1.33}$$

In the time delay domain, the inverse Fourier transform of $\Phi_H(\Delta f)$ is called the ***Power Delay Profile*** (**PDP** of the channel and gives the average power of the multi-path components as a function of time delays. It is given by [2]

$$P(\tau) = \int_{-\infty}^{+\infty} \Phi_H(\Delta f)e^{j2\pi\Delta f\tau}d\Delta f = \sum_p \sigma_p^2\delta(\tau_p-\tau_q) \tag{1.34}$$

The maximum delay τ_{max} or the standard deviation σ_τ of the **power delay profile** are often used to measure the delay spread $\Delta\tau$ [2, 32]. The standard deviation σ_τ is obtained as The Mobile Radio Channel

$$\sigma_\tau = \frac{1}{\Phi_H(0)}\int_0^{+\infty}(\tau-\tau_m)^2P(\tau)d\tau \tag{1.35}$$

where τ_m denotes the mean of the power delay profile.

Doppler Spread The Doppler spread is the dual of the delay spread. It describes the shift in frequency of the signal spectrum caused by the motion of the objects within the propagation environment. The different ***Doppler shifts*** ν_p of the different paths make the transfer function $H(t,f)$ vary in (3.11). The attenuation experienced by the signal

spectrum $S(f)$ at frequency f is therefore time variant and this phenomenon is referred to as time selectivity. The severity of the time selectivity is measured by the product of the *Doppler frequency* f_d with the time span the receiver needs to process the incoming signal. If coherent detection is assumed, where each data symbol is processed independently, the processing time is the symbol length T_s. So, if the Doppler frequency f_d is much lower than the processing rate $\frac{1}{T_s}$, the transfer function $H(t, f)$ stays almost constant within the symbol time T_s, and the channel is said to be *slow*. In contrast, if the *Doppler frequency* is larger than the processing rate, then the transfer function varies within the processing time, and the channel is called to be *fast* . By analogy to the spaced frequency correlation function $\Phi_H(\Delta f$ for frequency selectivity, the spaced time correlation function $\Gamma_H(\Delta t)$ reflects the time selectivity. It measures the correlation between $H(t, f)$ at different time instants and is defined as [2]

$$\Gamma_H(\Delta t) = \frac{1}{2}E[H(t, f)H^*(t - \Delta t, f)] \tag{1.36}$$

The fact that the autocorrelation function $\Gamma_H(\Delta t)$ only depends on time difference t results from the assumption that the transfer function $H(t, f)$ is a Wide Sense Stationary (WSS) process. Under this assumption, the scatterers at different Doppler shifts $\{\nu_p\}$ are uncorrelated.
From the time autocorrelation function $\Gamma_H(\Delta t)$, the so-called ***Doppler power spectrum*** is derived by Fourier transform, which yields

$$S_H(\nu) = \int_{-\infty}^{+\infty} \Gamma_H(\Delta t)e^{-j2\pi\nu\Delta t}d\Delta t \tag{1.37}$$

As for the coherence bandwidth, a measure called the coherence time is determined here from the time autocorrelation function in order to indicate the time span during which the transfer function $H(t, f)$ roughly stays constant. Again the definition of the coherence time is somewhat subjective and depends on the form of the Doppler power spectrum.

1.6 Concluding Remarks

In this chapter, we reviewed some important concepts from signals and systems theory. We also reviewed standard adaptive filtering techniques: LMS and RLS. We then introduced the different parameters for Bayesian adaptive filtering which are used in the remining of the thesis where we give the definition of each parameter followed by an illustrative example.

Since this thesis is particularly concerned with, although not limited to, parameter estimation of linear systems in state-space form the properties of such systems formed a particular focus of this chapter.

Chapter 2

Optimal State Estimation

Kalman smoothers and Kalman filters form an important component of algorithms developed later in this thesis. While most of the results presented in this chapter are well-known to much of the adaptive filtering research, some results, such as the recursion equation and the fixed interval Kalman smoother (Lemma 7), are either new or obscure in origin. Even in the case of the better-known material, its importance to this thesis merits its restatement. Following an introductory section on notation, we present a derivation of a recursive Kalman filter and include a practical strategy for ensuring that the results of the filter are numerically stable. An identical approach is then taken for the Kalman smoother. Our attention is then directed toward the stability of the time-varying Kalman filter and smoother. Recognizing that time-varying systems form the primary focus of this thesis, we first specialize to that case before launching a discussion on this topic. This approach has the advantage of allowing the presentation to be far more straightforward and streamlined. The simplicity of the smoother presented in Section 8 allows some well-known results demonstrating the stability of the filter to be extended, in a series of new results, to the case of the smoother.

2.1 Kalman Filter

The Kalman filter [148] solves the problem of finding filtered or predicted estimates of the state of time-varying linear state space systems, described by equations 6.1, from input-output data. The filter is distinguished by the fact that, for this class of systems, it is the overall minimum mean-square error estimator [149]. Since the model structure postulated in equations 6.1 involves Normally distributed random variables the state estimation problem is solved by computing the expected value of the state sequence conditional upon the available data (see Theorem 1 of [148]). Significantly, the Markovian nature of the underlying system allows this task to be achieved using a simple set of recursive equations. Derivations of the Kalman filter are widely available in the literature [150, 151, 152, 153] but many of these consider only simple time-series models. The results in this thesis require a derivation involving exogenous inputs. Such derivations are more involved than for the pure time-series case and appear surprisingly infrequently in the control literature. The proof of Lemma 7 (Kalman Filter), which is based upon the approach taken by Kalman [149], is included here for reasons of completeness. We begin by introducing a lemma about orthogonal projection, which can be found, for example, in Doob [154]. The proof presented here is based on [151].

Lemma 5 ***(Orthogonal Projection Lemma)****. Let X be a normal space, $x \in X$, and let Y be a subspace of X. Then $\hat{x} \in Y$ satisfies*

$$m \in_{\alpha \in Y} \|x - \alpha\|^2 = \|x - \hat{x}\|^2,$$

if, and only if,

$$< x - \hat{x}, \alpha >= 0,$$

for all $\alpha \in Y$,

Proof 6 *The "if" part: Suppose that $< x - \hat{x}, Y >= 0$. Take any $\alpha \in Y$, $\alpha \neq 0$. Then*

$$\begin{aligned} < x - \hat{x} + \alpha, x - \hat{x} + \alpha > &= < x - \hat{x}, x - \hat{x} > + 2 < x - \hat{x}, \alpha > + < \alpha, \alpha > \\ &= < x - \hat{x}, x - \hat{x} > + < \alpha, \alpha > \\ &> < x - \hat{x}, x - \hat{x} > . \end{aligned}$$

Now for the *only if* part: Suppose that there exists an α such that

$$< x - \hat{x}, \alpha >= \beta \neq 0$$

Then, for any scalar λ,

$$< x - \hat{x} + \lambda\alpha, x - \hat{x} + \lambda\alpha >= \|x - \hat{x}\|^2 + 2\lambda\beta + \lambda^2\|\alpha\|^2.$$

Then, for $\lambda = -\frac{\beta}{\|\alpha\|^2}$,

$$\|x - \hat{x} + \lambda\alpha\|^2 = \|x - \hat{x}\|^2 - \frac{2\beta^2}{\|\alpha\|^2} + \frac{\beta^2}{\|\alpha\|^2} < \|x - \hat{x}\|^2.$$

2.2 State-space signal model

The Kalman filter is an application of the general results of sequential estimation.
Let us consider the system defined below

$$\begin{aligned} H_k &= AH_{k-1} + W_k \\ y_k &= X_k^H H_{k-1} + v_k \end{aligned} \tag{2.1}$$

The Kalman filtering problem can be stated as follows: given a sequence of measurements $y_0, y_1, y_2, \ldots$, determine a sequence of estimates of the state of the system H_k in a computationally feasible, recursive manner.

2.3 Kalman filter: The Bayes approach

In this section we derive the Kalman filter from the Bayesian point of view. For the Bayesian approach, we assume that the noise processes are Gaussian distributed. Then the Bayes estimate of H_k amounts to finding the conditioned mean of H_k, given the observations.
The key equation in the Bayes derivation is the time update step,

$$f(H_k \mid Y_k) = \int f(H_k \mid H_{k-1}) f(H_{k-1} \mid Y_k) dH_{k-1} \tag{2.2}$$

from which the estimate is propagated using the state update equation into the future; and

$$\begin{aligned} \underbrace{f(H_k \mid Y_{k+1})}_{posterior} &= \frac{f(y_{k+1} \mid H_k)}{f(y_{k+1} \mid Y_k)} \underbrace{f(H_k \mid Y_k)}_{prior} \\ Y_k &= [y_k, \ldots, y_{k-M}] \end{aligned} \tag{2.3}$$

which is the measurement update step.
We will begin by finding explicit formulas for the time-update in (2.2).

1. The density $f(H_{k-1} \mid Y_k)$ corresponds to the estimate of H_{k-1}, given the measurements up to time k. Under the assumption and using the notation just introduced, the random variable H_{k-1} conditioned upon Y_k is Gaussian,

$$H_{k-1} \mid Y_k \sim N(\hat{H}_{k-1|k-1}, P_{k-1|k-1}) \tag{2.4}$$

2. The density $f(H_k \mid H_{k-1})$ is obtained by noting from (6.1) that, conditioned upon H_{k-1}, H_k is distributed as

$$H_k \mid H_{k-1} \sim N(AH_{k-1}, Q) \tag{2.5}$$

Inserting (2.4) and (2.5) into (2.2) and performing the integration (which involves expanding and completing the square), we find that $H_k \mid Y_k$ is Gaussian, with mean

$$\hat{H}_{k|k-1} = A\hat{H}_{k-1|k-1} \tag{2.6}$$

and the error covariance is given by

$$P_{k|k-1} = AP_{k-1|k-1}A^H + Q \tag{2.7}$$

Equation (2.6) provides a means to propagate the estimate ahead in time, in the absence of measurements, and (2.7) shows that, without measurements, the estimate covariance grows in time.

Let us now examine the update step in (2.4). This is a Bayes update of a Gaussian random variable. The mean of $H_k \mid Y_k$ is obtained using a Bayesian technique, in which the mean of prior is updated:

$$\begin{aligned} \hat{H}_{k|K} &= E[H_k \mid Y_k] = \hat{H}_{k|k-1} + R_{hy,Y_k} R^{-1}_{yy,Y_k} \\ &\quad \times (y_k - E[y_k \mid Y_k]). \end{aligned} \tag{2.8}$$

We now examine each component of this mean value:

1. Let R_{hy,Y_k} denote the correlation, conditioned upon Y_k:

$$R_{hy,Y_k} = E[(H_k - E[H_k])(y_k - E[y_k])^H \mid Y_k]. \tag{2.9}$$

Then we have

$$\begin{aligned} R_{hy,Y_k} &= E[(H_k - \hat{H}_{k|k-1}])(X_k^H (H_k - \hat{H}_{k|k-1}) + v_k)^H \mid Y_k] \\ &= P_{k|k-1} X_k \end{aligned} \tag{2.10}$$

2. Let R_{yy,Y_k} denote the covariance of y_k, conditioned upon Y_k:

$$\begin{aligned} R_{yy,Y_k} &= E[(y_k - E[y_k])(y_k - E[y_k])^H \mid Y_k] \\ &= E[(X_k^H (H_k - \hat{H}_{k|k-1}) + v_k) \\ &\quad \times (X_k^H (H_k - \hat{H}_{k|k-1}) + v_k)^H \mid Y_k] \\ &= X_k^H P_{k|k-1} X_k + \sigma_v^2 \end{aligned} \tag{2.11}$$

3. The mean $E[y_k \mid Y_k]$ is equal to $X_k^H \hat{H}_{k|k-1}$.

Putting the three expression together, we obtain the following update step:

$$\hat{H}_{k|k} = \hat{H}_{k|k-1} + P_{k|k-1} X_k (X_k^H P_{k|k-1} X_k + \sigma_v^2)^{-1} (y_k - X_k^H \hat{H}_{k|k-1}) \tag{2.12}$$

It will be convenient to let

$$K_k^f = P_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1} \tag{2.13}$$

so that the mean update can be written as

$$\hat{H}_{k|k} = \hat{H}_{k|k-1} + K_k^f(y_k - X_k^H \hat{H}_{k|k-1}). \tag{2.14}$$

The quantity K_k^f is called the *Kalman gain*.

Let us now consider the covariance of $H_k \mid Y_k$, which is the variance of the estimator error $\tilde{H}_{k|k} = H_k - \hat{H}_{k|k}$. In that case we found that the conditional density $X \mid Y$ had covariance

$$COV(X \mid Y) = R_{xx} - R_{xy}R_{yy}^{-1}R_{yx}.$$

To apply this formula, we identify the random variable X with $\hat{H}_{k|Y_k}$, and the observation Y with the observation y_k. The covariance R_{xx} is thus analogous to $P_{k|k-1}$. The matrix R_{xy} is analogous to $R_{xy,Y}$ and R_{yy} is analogous to $R_{yy,Y}$. Therefore we have

$$\begin{aligned} P_{k|k} &= P_{k|k-1} - P_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}X_k^H P_{k|k-1} \\ &= (I - K_k^f X_k^H)P_{k|k-1}. \end{aligned} \tag{2.15}$$

Lemma 7 ***Kalman Filter.*** *The minimum variance (Kalman) filter for the system (6.1) is given by the following recursion.*

$$\begin{aligned} \hat{H}_{k|k-1} &= A\hat{H}_{k-1|k-1}, \\ P_{k|k-1} &= AP_{k-1|k-1}A^H + Q, \\ K_k^f &= P_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}, \\ \hat{H}_{k|k} &= \hat{H}_{k|k-1} + K_k^f(y_k - X_k^H \hat{H}_{k|k-1}), \\ P_{k|k} &= (I - K_k^f X_k^H)P_{k|k-1}. \end{aligned} \tag{2.16}$$

2.4 Fixed-Point Kalman Smoothing

A fixed-point Kalman smoother [45] is one which calculates a sequence of estimates of the state at some predetermined sampling instant, k . The key idea is that the estimate of the state vector, $\hat{H}_{k|m}$, is recursively refined as the amount of data increases (that is, as m increases). This is in contrast to the necessarily one fixed-interval Kalman smoother which operates upon a block (or fixed-interval) of data, Y_m, and where a single state estimate, $\{H_{k|m}\}_{k=1}^{m}$, is calculated for each value of $k = 1, 2, \ldots, m$. Significantly, each fixed-point Kalman smoother may be implemented via a Kalman filter.

Lemma 8 ***Fixed-Point Kalman Smoothing*** *The fixed-point smoothed estimate of H_k given m data points, $\hat{H}_{k|m}$, and its associated error covariance matrix $P_{k|m}$, for the model structure (6.1) may be calculated by applying the Kalman filter recursions*

$$\begin{aligned} P_{k|k} &= P_{k|k-1} - P_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}X_k^H P_{k|k-1} \\ &= (I - K_k^f X_k^H)P_{k|k-1}. \end{aligned} \tag{2.17}$$

The minimum variance (Kalman) filter for the system (6.1) is given by the following recursion.

$$\begin{aligned} \hat{H}_{k+1|k} &= A\hat{H}_{k|k-1} + AK_k^f(y_k - X_k^H\hat{H}_{k|k-1}), \\ \hat{H}_{k|k} &= \hat{H}_{k|k-1} + K_k^{'}(y_k - X_k^H\hat{H}_{k|k-1}), \\ K_k^f &= P_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}, \\ K_k^{'} &= P_{k|k-1}^{'}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}, \\ P_{k+1|k} &= AP_{k|k-1}A^H + Q - AP_{k|k-1}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}X_k P_{k|k-1}A^H, \\ P_{k+1|k}^{'} &= P_{k|k-1}^{'}A^H - P_{k|k-1}^{'}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}X_k P_{k|k-1}A^H, \\ P_{k|k} &= P_{k|k-1} - P_{k|k-1}^{'}X_k(X_k^H P_{k|k-1}X_k + \sigma_v^2)^{-1}X_k P_{k|k-1}^{'}. \end{aligned} \tag{2.18}$$

Proof. See [45]

2.5 Discussion

The main objective of this chapter was to present a theoretical background to the problem of optimal state estimation so as to lay a foundation for developments in later chapters. For the model structures considered in this thesis, the problem of state estimation is solved, under differing assumptions upon the availability of data, by the Kalman smoother or the Kalman filter. The main difference is that the Kalman filter uses data only up to the present, whereas the Kalman smoother uses both past and future data in its calculations and therefore is of considerable interest in ofline settings. Interestingly, it turns out that a set of Kalman smoothed state estimates may be calculated by using a sequence of Kalman filters. Numerically robust versions of both the filter and smoother were derived. An advantage of this smoothing scheme is that the covariance matrices are calculated in such a manner as to ensure that they are positive semi-definite. Furthermore, it is more straightforward and simple than pre-existing alternatives (see, for example, the treatment of RTS Kalman smoothers in Kailath et al.[43]).

Chapter 3

Overview of Parameter Estimation

3.1 Introduction

The science of System Identification can be broadly defined as the theory and practice of deriving models from experimental data. While the statistical fundamentals of this field have in many cases been in existence for almost a century it is really only since the 1960s that work in this area has been undertaken intensively. Since that time the concerted effort of many researchers has done a great deal to unify what initially seemed to be a collection of disparate ad-hoc approaches. Arguably, this process of theoretical consolidation allowed the subject to experience an apotheosis as a mature research area with the publication of a number of substantial texts on the topic (for example, [46] and [48]) and together with a sophisticated and usable software package [47]. The problem of system identification can be divided into a number of subproblems. These are stated below:

- Experiment design,
- Data collection,
- Selection of model structure,
- Selection of model,

- Model validation.

This thesis is concerned primarily with the problem of selecting a model from a set of candidates. In the ensuing text we shall assume that the experiment design, data collection and model structure selection operations have already been performed. In addition, we shall assume that the model structure is a parametric one (that is the model structure is parameterised by a finite-dimensional vector of real numbers, generally denoted θ). The task of model selection, therefore, is equivalent to that of selecting a suitable parameter vector from a set of candidates. This practice is known as " Parameter Estimation" .
We are interested in estimating the parameters of a system and have been given a set of data,

$$Z \triangleq (H_k, Y_k) \tag{3.1}$$

consisting of a sequence of discretely sampled measurements of its inputs and outputs. A parameter estimation method is a mapping from the data Z to an element of a set of candidate parameter vectors, denoted Θ. That is,

$$Z \rightarrow \hat{\theta}(Z) \in \Theta. \tag{3.2}$$

The symbol $\hat{\theta}(Z)$

in equation (8.4) is the estimate based upon the pairs of input-output data Z. Furthermore, when considering iterative estimators we shall denote the estimate at the $k-th$ iteration, based upon the data set Z as $\hat{\theta_k}(Z)$ or, using a more relaxed notation, $\hat{\theta_k}$, where dependence upon ZN is tacitly assumed. What is not immediately apparent in this formulation is that an important element of many parameter estimation schemes is a criterion (or cost) function, $V_N(\theta)$. The purpose of such a function (which, strictly speaking, is also a function of the data, Z) is to define the exact manner in which the mapping (8.4) occurs. The basic idea is to choose as the estimate an element of Θ that makes $V_N(\theta)$ small. That is, the parameter estimate is typically computed according to the relationship

$$\hat{\theta} = arg \min_{\theta \in \Theta} V_N(\theta) \tag{3.3}$$

Clearly, the criterion function can be a major influence upon the properties of a parameter estimator. In the remainder of this chapter we present a number of common approaches to parameter estimation in a statistical framework and discuss their properties and common elements. Specifically, we shall discuss the Maximum Likelihood and Prediction-Error parameter estimation methods [46]. We shall also consider a Subspace-based parameter estimation algorithm. The purpose of this material is to provide a background to this topic and some context into which the algorithms developed in later chapters can be fitted. In keeping with most related work [46, 49] we shall focus particularly upon the asymptotic properties of the consistency and relative efficiency of these estimation schemes. We begin though, by discussing these properties.

3.2 Maximum Likelihood Methods

The Maximum Likelihood (ML) approach to parameter estimation is very well-established, rooted in seminal work of Fisher [51] in the early twentieth century. Since that time the method has been investigated under a wide variety of modeling assumptions [150, 52] and its properties are very well understood. According to this approach one embraces a probabilistic framework in order to treat the sequence of observations as a realization of a stochastic process. On the basis of this data, one attempts to estimate a parameter vector so that the likelihood of having seen such a realization is maximized - hence the name. Suppose that the joint (conditional) probability density function of a set of observations, Z, is known to be $p_\theta(Z)$, where $\theta \in \Theta$ parameterizes the probability density function and that is the set of allowable parameter vectors. Now, if one observes the realization $Z = Z^*$then a ML estimator for those parameters is

$$\hat{\theta}(Z^*) = arg\max_\theta p_\theta(Z) \mid_{Z=Z^*} \tag{3.4}$$

Once a series of measurements, Z, has been taken - that is Z fixed to be some Z^*, the function

$$p_\theta(Z) \mid_{Z=Z^*}, \tag{3.5}$$

becomes a deterministic function of θ. The resulting maximization operation contains

no stochastic components. Instead of performing the maximization operation in equation (6.17) it is more common to recognize that, since the logarithm function is monotonically increasing and the mathematical operation of maximizing a function is the same as minimizing its negative, it is entirely equivalent to select the estimate as the element that minimizes the negative logarithm of $p_\theta(Z)$. That is, we can determine an estimate by solving

$$\hat{\theta} = arg\min_{\theta}(-L(\theta)) \tag{3.6}$$

where the $L(\theta)$ is the log-likelihood function defined as

$$L(\theta) \rightarrow logp_\theta(Z) \mid_{Z=Z^*}, \tag{3.7}$$

instead of (6.17). Comparing the basic approach of equation (6.2) to that of equation (3.6) reveals that, in this formulation, the maximum likelihood criterion function is actually the negative log-likelihood function. That is,

$$V(\theta) \rightarrow -L(\theta). \tag{3.8}$$

The approach of minimizing the negative log-likelihood function is particularly popular when the underlying probability density functions are exponential. For example, when $p_\theta(Z)$ is a Gaussian probability density function [151] and the data are independent, the resulting negative log-likelihood function appears in an attractive form largely consisting of a sum of quadratic terms. One of the most valuable aspects of the ML method is that it provides a general framework for solving a wide range of parameter estimation problems provided that the distribution of the data is known. For example, in the early stages of the identification experiment design phase the ML method provides a clear rationale for selecting one particular criterion function over another.

3.2.1 Properties

In this section we discuss briefly some well-known and attractive properties of the maximum likelihood method. First, we present a theorem that outlines the startling fact that

the best (in the sense of meeting the Cramer-Rao bound) unbiased estimators are always Maximum Likelihood estimators.

Property 9 *Whenever there exists an unbiased estimator which achieves the Cramer-Rao lower bound then it is also the maximum likelihood estimator. Proof. See [55]. Second, ML estimates are strongly consistent in simple i.i.d situations and achieve the Cramer-Rao bound:*

Property 10 *Let $\hat{\theta}(Z)$ designate a maximum likelihood estimator of θ^* based upon m i.i.d. random variables, Z, then*

$$\hat{\theta}(Z) \rightarrow \theta^* \tag{3.9}$$

$$\sqrt{m}(\hat{\theta}(Z) - \theta^*) \rightarrow N(0, \Gamma^{-1}) \tag{3.10}$$

where Γ is the average value of Fisher s information matrix per sample. Proof. See [56] and [57].

These properties, easily established in the i.i.d. case, have been extended to many more general cases [58, 49]. While the unbiased maximum likelihood estimators have good asymptotic properties (as assessed by the covariance of the resulting estimates), it should be realized that there are other ways of measuring the performance of estimators. Ljung [46] notes that the small sample behaviour of maximum likelihood estimators has sometimes been criticized for being poor. In addition, for many problems it is the exception rather than the rule that an unbiased ML estimator can be found [55].

3.2.2 Implementation

The difficulty associated with implementing the ML method lies in performing the $arg\min_\theta$ operation of equation (3.6). Since the likelihood function is generally a non-convex function of its parameters,θ , the search for the minimizer must be undertaken iteratively. One

commonly-adopted strategy is to use an iterative gradient-based search scheme such as the well-known Newton method or perhaps one of its derivatives [59, 48]. For example, a (damped) Newton method solution of a Maximum Likelihood problem employs iterations of the form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \mu_k[H(\hat{\theta}_k)]^{-1}J(\hat{\theta}_k) \tag{3.11}$$

in order to update the current estimate $\hat{\theta}_k$ to a better estimate $\hat{\theta}_{k+1}$ (that is, one associated with a higher likelihood). Here μ_k is a user-chosen step length, $H(\hat{\theta}_k)$ is the Hessian of $L(\theta)$ at $\hat{\theta}_k$, defined as

$$H(\theta_k) = \frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta) \mid_{\theta=\hat{\theta}} \tag{3.12}$$

and $J(\hat{\theta}_k)$ the gradient of $L(\theta)$ at $\hat{\theta}_k$

$$J(\hat{\theta}_k) = \frac{\partial}{\partial\theta}L(\theta) \mid_{\theta=\hat{\theta}} \tag{3.13}$$

Gradient-based approaches, as exemplified by equation (3.12), implicitly exploit the continuity of the likelihood function, $L(\theta)$, in order to calculate the required first and second (partial) derivatives. An advantage of using this family of optimisation algorithms is that, under appropriate regularity conditions, the rate of convergence can be quite fast. For example, Newton's method can converge quadratically [59]. On the other hand, the necessary gradient calculations (3.12), (3.13) mandate the use of a specific model parameterisation and, for multivariable systems, this is a notoriously difficult problem [60, 43]. On the other hand, overparameterised model structures, that is those with nonminimal parameterisations, necessarily lead to singular Hessian matrices (defined by equation (3.12)) and to complications when evaluating the parameter update equation (3.11). The question of choosing a parameterisation for the model defined in (3.1) chapter 3, so that the resulting search is well-conditioned is still an open research topic [44, 45], and is a prime motivator for further developments in this thesis. For the special case of ML problems an exciting alternative to the gradient-based methods is provided by the Expectation Maximisation (EM) algorithm [101]. This approach, which arose in the mathematical statistics literature, is one that has received relatively little attention in the control community. This algorithm is comprehensively detailed in the next chapter.

3.3 Prediction Error Methods

The idea of comparing a model s predicted output to a measured output is a long-standing one in system identification [87, 88] and the term Prediction-Error Methods (PEM) was developed to unify a number of seemingly disparate approaches in a manner closely related to the ML method [76]. As their name suggests, the prediction-error methods provide a set of estimators that evaluate candidate models by how well they predict the system output. That is, their criterion function, $V(\theta)$, is a function of prediction errors.
According to this estimation framework one interprets a model class as a mapping from the space of allowable parameters, Θ , and the measured data, Z, to a causal prediction of the output. The one-step ahead model-based prediction of the system output, $\hat{y}_k(\theta)$, can then be written as a function of past values of $\{y_k\}$, past and current values of H_k, and the parameter vector θ as

$$\hat{y}_k(\theta) = g(\theta, Y_{k-1}, H_k) \tag{3.14}$$

Of course, the exact nature of the function $g(.)$ is determined by the model structure being used. A crucial element of the prediction error approach is that the criterion function, $V(\theta)$), is chosen to depend upon the prediction errors - that is, the difference between the system output $\{y_k\}$and the model-based prediction, $\hat{y}_k(\theta)$ given by equation (3.14). One common prediction error criterion function is the following simple weighted quadratic one

$$V(\theta) = \frac{1}{m} \sum_{k=1}^{m} \epsilon_k^T(\theta) \wedge^{-1} \epsilon_k(\theta) \tag{3.15}$$

where

$$\epsilon_k(\theta) = y_k - \hat{y}_k(\theta) \tag{3.16}$$

and is a symmetric positive definite matrix that serves to weight the various components of ϵ_k according to their relative importance in the criterion function. Of course, the estimate $V(\theta)$ is still formed via equation (6.17). Since this cost function is not only a

weighted least-squares one and but also happens to be the same as that employed by the ML method, we may conclude that under certain circumstances the three estimation techniques coincide.
One reason for the popularity of the simple criterion function embodied by (3.15) is that the problem of finding solutions to least-squares problems is very well studied [59]. Indeed, in some circumstances, it is possible to minimise equation (3.15) in closed-form.

In a more general prediction error setting though, one may define a (possibly) time-varying, weighted norm-like function, $l(k, \epsilon_k)$, and then choose $\hat{\theta}(Z)$ as the global minimiser of the cost function

$$V(\theta) = \frac{1}{m}\sum_{k=1}^{m} l(k, \epsilon_k). \tag{3.17}$$

3.3.1 Properties

There is an enormous amount of literature pertaining directly or indirectly to the properties of the prediction error methods [61, 49]. Much of this interest is due to the fact that under common assumptions prediction error techniques are equivalent to least-square methods [62, 153] or to the Maximum Likelihood approaches [105, 106]. It turns out that, under mild conditions upon the system, the set of models, the criterion function and input signal, the prediction error methods satisfy the following asymptotic properties due to [49] and [46].

Property 11 ***(Strong Consistency)****. Let Θ be a set of parameters and $\hat{\theta}(Z)$ and $V(\theta)$ be defined by equations (6.17) and (3.17), respectively. Then, under the appropriate regularity conditions,*

$$\hat{\theta}(Z) \rightarrow_{a.s} \theta^* \, as \, m \rightarrow \infty, \tag{3.18}$$

(3.19)

where

$$\theta^* = arg\min_{\theta\in\Theta} li{-}_{m\rightarrow\infty} E\{V_m(\theta)\}. \tag{3.20}$$

Note that if the model structure is sufficiently flexible then multiple solutions of equation (3.20) may exist and then property above must be amended to state that the estimate, $\hat{\theta}(Z)$, converges to a set of cost function minimisers.

Property 12 ***(Asymptotic Normality).*** *Let $\hat{\theta}(Z)$ and θ^* be defined by equations (6.17) and (3.20), respectively. Then, under the appropriate regularity conditions, there exists a sequence of positive semi-definite matrices $\{P_m\}$ such that*

$$\sqrt{m}P_m^{\frac{-1}{2}}(\hat{\theta}(Z) - \theta^*) \rightarrow N(0, I)\, as\, m \rightarrow \infty. \tag{3.21}$$

$$\tag{3.22}$$

Clearly, an estimate will not exhibit asymptotic normality unless θ^* is the unique solution of equation (3.20). For further information on the asymptotic properties of the PEMs we point the interested reader to [46] or, for a treatment of greater depth and generality, [49] and [46].

3.3.2 Implementation

Properties (11) and (12) were both derived under the assumption that equation (6.17) is globally solvable. However, for many scenarios, the cost function $V(\theta)$ is non-convex in θ and therefore a closed-form solution of equation (6.17) is not available. Again, gradient based search algorithms are a popular choice for solving the resulting optimisation problem Eq. (6.17)) and therefore the implementational advantages and disadvantages of the prediction-error methods bear a striking similarity to those of the ML approach. Of course, for the PEM case the Hessian matrix and gradient vector appearing in equation ((3.11) are defined to be

$$H(\theta_k) = \frac{\partial^2}{\partial\theta\partial\theta^T}V(\theta)\mid_{\theta=\hat{\theta}} \tag{3.23}$$

and

$$J(\hat{\theta}_k) = \frac{\partial}{\partial \theta} V(\theta) \mid_{\theta=\hat{\theta}} . \tag{3.24}$$

3.4 Subspace-based Parameter Estimation Methods

The study of State Space Subspace-based System Identification (4SID) methods [63, 91, 92] has been of enormous recent interest. Belying the recent nature of this activity, the ideas involved actually go back many years, at least to Akaike[87] whose approach targeted the types of stochastic estimation problems considered in this thesis. There are too many varieties of 4SID algorithms to detail all of them here but the basic unifying theme of the time-domain versions is the extraction of estimates of system state-space matrices directly from data by first dividing that data into past and future data and then projecting the future data onto the space spanned by the past data. The technique of projecting onto subspaces in some ways tends to tie the resulting algorithms more closely to linear system theory than the other methods described in this chapter. That is not to say that subspace algorithms have not been applied successfully to problems of non-standard linear or nonlinear system identification. There have been notable examples of their use on errors-in-variables [110], bilinear [107, 108] and linear parameter-varying [109] systems. However many of these extended subspace algorithms are probably not applicable to large nonlinear systems with the current level of computing technology, since the resulting data matrices tend to grow at an exponential rate with increasing model order. The benefits of using these methods, particularly in the linear case, are considerable - parameter estimates may be extracted non-iteratively directly into state-space form, thus making them ideal for multivariable identification. Furthermore, without the need for gradient-based search there is no necessity for explicitly parameterising the state-space matrices - the resulting estimates are fully parameterised. Other advantages of these approaches lies in their numerical simplicity and the reliability of their implementation. The key operation required is one of projection which may be performed with Singular Value Decomposition or even QR factorisation. Unfortunately, the SVD operation in particular makes these algorithms non-linear in the data and this renders difficult any analysis of the statistical performance of the approach. In spite of the limitations imposed by their nonlinear nature,

great strides have been made in the analysis of various subspace algorithms. Analytical studies of the consistency and relative efficiency [114], and asymptotic normality [113] of subspace estimates been conducted. Moreover, simulation studies in, for example [111] and [114], tend to suggest that the relative efficiency of subspace algorithms is close to that of maximum likelihood algorithms. At present, there is only one known case in which subspace algorithms achieve the Cramer-Rao bound, and that is when a CCA algorithm is used and the input is white [112]. Despite all of this, some work remains to be done in this area so that the depth and richness of theory enjoyed by users of maximum likelihood and prediction-error methods is also enjoyed by users of subspace-based parameter estimation methods.

3.5 Conclusions

In this chapter we presented a number of popular parameter estimation schemes and described some of their properties.

The prediction error and maximum likelihood methods are both well-studied approaches to estimation and benefit from a correspondingly deep set of supporting theory. Difficulties with these algorithms can arise upon their implementation with common gradient-based techniques requiring an explicit model parameterisation. A key observation here was that the process of extending such optimisation algorithms from the Single-Input, Single-Output (SISO) case to that of multivariable systems is not at all straightforward. Indeed, this process can be quite difficult.

A popular and numerically robust alternative to these algorithms is provided by the subspace-based estimation methods. These cope admirably with multivariable parameter estimation as they naturally employ compact and attractive state-space model structures. On the other hand, the theory so-far developed for these algorithms lags that of the PE and ML techniques. In the remainder of this thesis we test the potential of the EM algorithm for solving parameter estimation problems. We begin, in the next chapter, with a thorough description of the EM algorithm.

Chapter 4

EM Algorithm

In this chapter, we introduce a means of maximum-likelihood estimation of parameters that is applicable in many cases when direct access to the necessary to make the estimates is impossible , or when some of the data is missing. Such inaccessible data are present, for example, when an outcome is a result of an accumulation of simpler outcomes, or when outcomes are clumped together (e.g., in a binning or histogram operation). There may also be data dropouts or clustering such that the number of underlying data points is unknown (censoring and/or truncation). The EM (expectation-maximization) algorithm is ideally suited to problems of this sort, in that it produces maximum-likelihood (ML) estimates of parameters when there is a many-to-one mapping from an underlying distribution to the distribution governing the observation. The EM algorithm consist of two primary steps: an expectation step, followed by maximization step. The expectation is obtained with respect to the unknown underlying variables, using the current estimate of the parameters and conditioned upon the observations. The maximization step then provides a new estimate of the parameters. These two step are iterated until convergence. The concept is illustrated in Fig. (5.1).

The EM algorithm was discorved and employed independently by several different research; see ([101]) brought their ideas together, proved convergence, and coined the term ”EM algorithm.” Since this seminal work, hundreds of papers employing the EM algorithm in many areas have been published. A typical application area of the EM algorithm is genetics, where the observed data is a function of the underlying, unobserved gene

patten; see, for example [115]. Another area is estimating parameters of mixture distributions, as in [116]. The EM algorithm has also been widely used in econometric, clinical, and sociological studies that have unknown factors affecting the outcomes [117]. Some applications to the theory of statistical methods are found in [118].
In the area of signal processing applications, the largest area of interest in the EM algorithm is maximum-likelihood tomographic reconstruction (see, for example, [119]). Another commonly cited applications is the training of hidden Markov models, especially for speech recognition, as in [120].

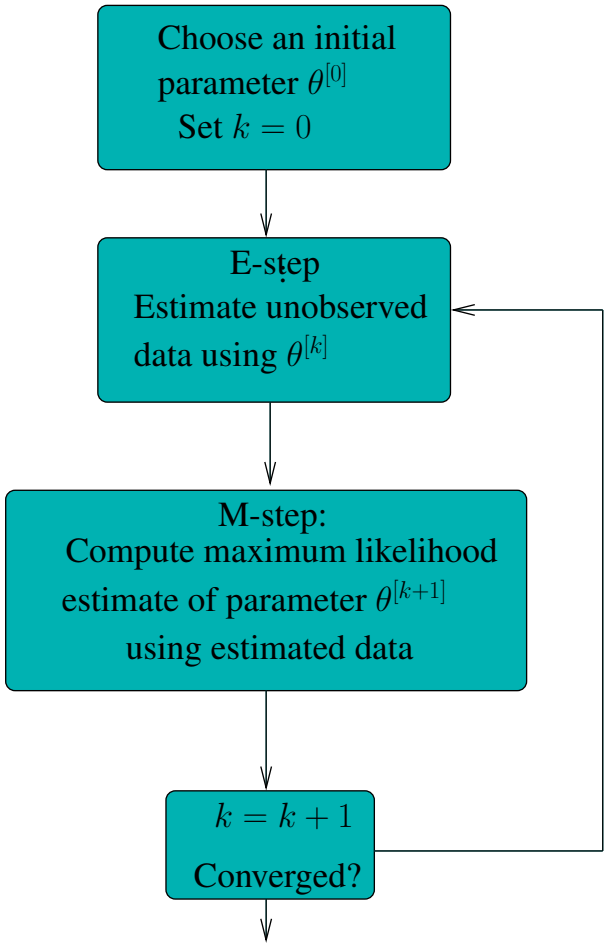

Figure 4.1: An overview of the EM algorithm. After initialization, the E-step and the M-step are alternated until the parameter estimate has converged (no more change in the estimate)

Other signal processing and engineering applications began appearing in the mid

$1980s$. These include: parameter estimation ([121, 122]), ARMA modeling [123], image modeling, reconstruction, and processing [124, 125], simultaneous detection and estimation [126, 127], pattern recognition and network training [128], direction finding [129], noise suppression [130], signal enhancement [131], spectroscopy, signal and sequence detection [132], time-delay estimation [133],and specialized development of the EM algorithm itself [134]. The EM algorithm also related to algorithms used in information theory to compute channel capacity and rate-distortion functions [135, 136], since the expectation step in the algorithm produces a result similar to entropy. The EM algorithm is philosophically similar to ML detection in the presence of unknown enraged with respect to the unknown quantity (i.e. the expected value likelihood function is computed) before detection, which is a maximization step (see, for example, [2], chapter 5).
The algorithm is presented in this thesis to estimate the model parameters in the time varying systems.

4.1 General Statement of the EM algorithm

Let Y denote the sample space of the observations, and $y \in^m R$ denote an observation from Y. Let Z denote the underlying space and let $z \in R^n$ be outcomes from Z, with $m < n$. The data z is referred to as the **complete data**. the complete data z are note observed directly, but only by means of y, where $y = y(z)$, and $y(z)$ is a many-to-one mapping. An observation y determines a subset of $\mathfrak{Z}$, which is denoted as $\mathfrak{Z}(y)$. Fig. (5.1) illustrates the mapping.
The pdf of the complete data is $f_Z(z \mid \theta)$, where $\theta \in \Theta$ is the set of parameters of the density. The pdf f is assumed to be continuous in θ and appropriately differentiable. The ML estimate of θ is assumed to lie within the region Θ. The pdf of the incomplete data is

$$g(y \mid \theta) = \int_{\mathfrak{X}(y)} f(z \mid \theta) dx. \tag{4.1}$$

Let

$$l_y(\theta) = g(y \mid \theta) \tag{4.2}$$

denote the likelihood function, and let

$$L_y(\theta) = log\, g(y \mid \theta) \tag{4.3}$$

denote the log-likelihood function.

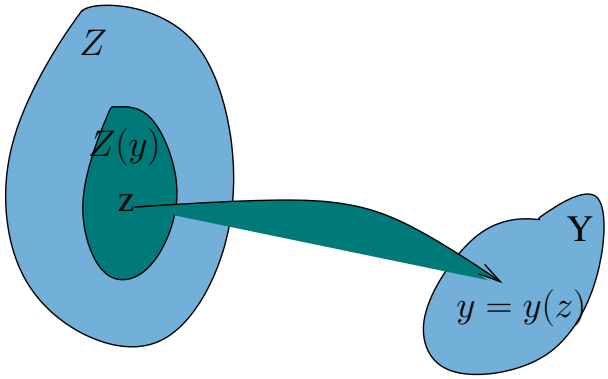

Figure 4.2: Illustration of many-to-one mapping from Z and Y. The point y is the image of z, and the set $Z(y)$ is the inverse map of y

The basic idea behind the EM algorithm is that we would like to find Θ to maximize $logf(z \mid \theta)$, but we do not have data z to compute the log-likelihood. So, instead, we maximize the expectation of $logf(z \mid \theta)$ given the data y and our current estimate of θ. This can be accomplished in two steps. Let $\theta^{[k]}$ be our estimate of the parameters at the $k-th$ iteration.

E-step.

$$Q(\theta \mid \theta^{[k]}) = E[logf(z \mid \theta) \mid y, \theta^{[k]}]. \tag{4.4}$$

It is important to distinguish between the first and second arguments of the Q functions. The second argument is a conditioning argument to the expectation and is regraded as fixed and known at every $E-step$. The first argument conditions the likelihood of the complete data.

M-step. Let $\theta^{[k+1]}$ be that value of θ that maximize

$$\theta^{[k+1]} = arg\max_{\theta} Q(\theta \mid \theta^{[k]}). \tag{4.5}$$

It is important to note that the maximization is with respect to the first argument of the Q function, the conditioner of the complete data likelihood.
The EM algorithm consists of choosing an initial $\theta^{[k]}$, then performing the $E - step$ and the $E - step$ successively until convergence. Convergence may be determined by observing when the parameters stop changing: for example, when $\|\theta^{[k+1]} - \theta^{[k]} < \epsilon\|$ for some ϵ and some appropriate distance measure $\|.\|$.

Example The general form of the EM algorithm as stated in (4.4) and (4.5)may be specialized and simplified somewhat by restrictions to distributions in the *exponential family*. These are pdfs of the form

$$f(z \mid \theta) = a(z)c(\theta)\exp[\pi(\theta)^T t(z)], \tag{4.6}$$

where θ is a vector of parameters for family, and where

$$t(z) = [t_1(z), \ldots, t_q(z)]^T \tag{4.7}$$

is the vector of sufficient statistics for θ. For exponential families, the $E - step$ can be written as

$$Q(\theta \mid \theta^{[k]}) = E[loga(z) \mid y, \theta^{[k]}] + \pi(\theta)^T E[t(z) \mid y, \theta^{[k]}] + logc(\theta) \tag{4.8}$$

Let $t^{[k+1]} = E[t(z) \mid y, \theta^{[k]}]$. Because a conditional expectation is an estimator, $t^{[k+1]}$ is an estimate of the sufficient statistic. In light of the fact that the $M - step$ will be maximizing

$$E[loga(z) \mid y, \theta^{[k]}] + \pi(\theta)^T t^{[k+1]} + logc(\theta) \tag{4.9}$$

with respect to θ and that $E[loga(z) \mid y, \theta^{[k]}]$ does not depend upon θ, it is sufficient to write the flowing.

E-step. Compute

$$t^{[k+1]} = E[t(z) \mid y, \theta^{[k]}]. \tag{4.10}$$

M-step. Compute

$$\theta^{[k+1]} = arg\max_{\theta} \pi(\theta)^T t^{[k+1]} + logc(\theta). \tag{4.11}$$

The EM algorithm has the advantage of being simple, at least in principle; actually computing the expectations and performing the maximization may be computationally taxing. Unlike other optimization techniques, it does not require the computation of gradients or Hessians, nor is it necessary to worry about setting set-size parameters, such as gradient descent algorithms.

4.1.1 Convergence of the EM Algorithm

For every iterative algorithm, the question of convergence must be addressed. Does the algorithm come finally to a solution, or does it iterate, ever learning but never coming to a knowledge of the truth? For the EM algorithm, convergence may be stated simply: at every iteration of the algorithm, a value of the parameter is computed so that the likelihood function of y does not decrease. That is, at every iteration, the estimated parameter provides an increase in the likelihood function (but will not decrease).
We present a proof of this general concept as follows. Let

$$k(z \mid y, \theta) = \frac{f(z \mid \theta)}{g(y \mid \theta)}. \tag{4.12}$$

and note that $k(z \mid y, \theta)$ may be interpreted as a conditional density. Then the log-likelihood function $L_y(\theta) = logg(y \mid \theta)$ may be written

$$L_y(\theta) = logg(z \mid \theta) - logk(z \mid y, \theta). \tag{4.13}$$

Define

$$H(\theta^{'} \mid \theta) = E[logk(z \mid y, \theta^{'}) \mid y, \theta].$$

Let M: $\theta^{[k]} \rightarrow \theta^{[k+1]}$ represent the mapping defined by the EM algorithm in (4.4) and (4.5), so that $\theta^{[k+1]} = M(\theta^{[k]})$.

Theorem 13 *$L_y(\theta^{[k+1]}) \geq L_y(\theta)$, with equality if and only if*

$$\begin{aligned} Q(M(\theta) \mid \theta) &= Q(\theta \mid \theta) \\ k(z \mid y, M(\theta)) &= k(z \mid y, \theta). \end{aligned}$$

That is, the likelihood function increases at each iteration of the EM algorithm, until the conditions for equality are satisfied and a fixed point of the iteration is reached. If θ^* is an ML parameter estimate, so that $L_y(\theta)^* \geq L_y(\theta)$ for all $\theta \in \Theta$, then $L_y M((\theta)^*) = L_y(\theta^*)$. In order words, ML estimates are fixed points of the EM algorithm. Since the likelihood function is bounded (for distributions of practical interest), the sequence of parameter estimates $\theta^{[0]}, \theta^{[1]}, \ldots, \theta^{[k]}$ yields a bounded nondecreasing sequence $L_y(\theta^{[0]}) \leq L_y(\theta^[1]) \leq \ldots L_y(\theta^{[k]})$, which must converge as $k \rightarrow \infty$.

Proof 14

$$L_y M(\theta) - L_y(\theta) = Q(M(\theta) \mid \theta) - Q(\theta \mid \theta) + H(\theta \mid \theta) - H(M(\theta) \mid \theta). \quad (4.14)$$

By the definition of the $M - step$, it must be the case that

$$Q(M(\theta) \mid \theta) \geq Q(\theta \mid \theta).$$

for every $\theta \in \Theta$. For any pair $(\theta', \theta) \in \Theta \times \Theta$, it is the case that

$$H(\theta' \mid \theta) \leq H(\theta \mid \theta).$$

This can be proven with Jensen's inequality, which states: If $f(z)$ is a concave function, then $E[f(z)] \leq f(E[x])$, with equality if and only if z is constant (nonrandom). This

inequality may be employed as follows.

$$\begin{aligned} H(\theta^{'} \mid \theta) - H(\theta \mid \theta) &= E[log\frac{k(z \mid y, \theta^{'})}{k(z \mid y, \theta)} \mid y, \theta] \\ &\leq logE[\frac{k(z \mid y, \theta^{'})}{k(z \mid y, \theta)} \mid y, \theta] \qquad (4.15) \\ &= log \int_{\mathfrak{Z}} \frac{k(z \mid y, \theta^{'})}{k(z \mid y, \theta)} k(z \mid y, \theta) dz \qquad (4.16) \\ &= log \int_{\mathfrak{Z}} k(z \mid y, \theta^{'}) dz \\ &= 0. \qquad (4.17) \end{aligned}$$

Equation (4.16) follows from Jensen's inequality, with $f(z) = log(z)$, which is concave; and (4.17) is true since $k(z \mid y, \theta)$ is a conditional density.
Examination of (4.14) in light of the $M - step$ and the conditions for equality in Jensen's inequality reveals that equality in the theorem can only hold for the stated conditions.
The theorem falls short of proving that the fixed point of the EM algorithm are fact ML estimates. The latter is true, under rather general conditions, but the proof is somewhat involved and is not presented here (see [99]).

Lemma 15 *Suppose that $\hat{\theta}_k$ is an instance of an EM algorithm such that*

1. *$\hat{\theta}$ converge to θ^**
2. *$\frac{\partial}{\partial \theta} Q(\theta, \hat{\theta}_k) = 0$*
3. *$\frac{\partial^2}{\partial \theta \partial \theta^T} Q(\theta, \hat{\theta}_k)$ is negative definite with eigenvalues bounded away from zero.*

Then

$$\frac{\partial}{\partial \theta} L(\theta) = 0, \qquad (4.18)$$

$$\frac{\partial^2}{\partial \theta \partial \theta^T} Q(\theta, \hat{\theta}_k) \quad \textit{is negative definite}$$

and

$$\frac{\partial}{\partial\theta}M(\theta) = [\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta,\theta^*)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}V(\theta,\theta^*) \tag{4.19}$$

Proof. See the appendix.
In order to see the utility of this lemma, note that if we linearise the EM algorithm about the point to which it is converging by finding its first-order Taylor expansion, then

we obtain

$$\begin{aligned}\hat{\theta}_{k+1} &= M(\hat{\theta}_k) \\ &\approx \theta^* + \frac{\partial}{\partial\theta}M(\theta)|_{\theta=\theta^*}(\hat{\theta}_k - \theta^*)\end{aligned}$$

and then

$$\tilde{\theta}_{k+1} \approx (\frac{\partial}{\partial\theta}M(\theta)|_{\theta=\theta^*})^{N-1}\tilde{\theta}_k \tag{4.20}$$

Equation (4.20) formulates the EM algorithm as an autonomous linear time-invariant system. Under such conditions it is well known that θ_k will converge to an optimal value at an exponential rate determined by the largest eigenvalue of $\frac{\partial}{\partial\theta}M(\theta)$.

In the next section we shall discuss in greater depth the rate of convergence of the EM algorithm in light of equation (4.21), and in particular how it can be affected by the choice of missing data.

4.1.2 The Role of Missing Data

The EM algorithm allows the user to choose what constitutes the missing data. One purpose of this data is to make the optimisation problem (4.11) easy to solve but one should recognise that this choice also has an important effect upon the speed of convergence of

the algorithm. Note that equation (4.19) may be re-expressed as

$$\begin{aligned}\frac{\partial}{\partial\theta}M(\theta) &= [\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta,\theta^*)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}V(\theta,\theta^*)\\ &= [\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta,\theta^*)]^{-1}\\ &\quad\times[\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta,\theta^*)-\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta)]\\ &= I-[\frac{\partial^2}{\partial\theta\partial\theta^T}Q(\theta,\theta^*)]^{-1}\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta)]\\ &= I-\Gamma_{aug}^{-1}\Gamma_{obs}\end{aligned} \tag{4.21}$$

by using equation (6.5)

$$\Gamma_{aug}=\frac{\partial^2}{\partial\theta\partial\theta^T}E_{\hat{\theta}}\{\log f_Z(z,\theta)\}|_{\theta=\theta^*} \tag{4.22}$$

is the expected information matrix of the complete data set and

$$\Gamma_{obs}=\frac{\partial^2}{\partial\theta\partial\theta^T}E_{\hat{\theta}}\{\log f_Y(Y,\theta)\}|_{\theta=\theta^*} \tag{4.23}$$

is the observed information matrix.

Note that the rate of convergence of the EM algorithm as shown by equation (4.20)is dictated by the largest eigenvalue of $\frac{\partial}{\partial\theta}M(\theta)$. If this eigenvalue has a magnitude close to unity, then the algorithm will be slow to converge. Conversely, fast convergence correspond to this eigenvalue being close to zero. Under this scenario, it follows from equation (4.21) that it is desirable to choose the missing data and filter coefficient sequence, so that the smallest eigenvalue of $\Gamma_{aug}^{-1}\Gamma_{obs}$ is as large as possible. Clearly, Γ_{obs} is independent of the missing data so therefore the key to ensuring fast convergence is to find a filter coefficient sequence so that Γ_{aug} is small.

4.2 Discussion

The EM algorithm provides a simple, iterative method for solving maximum likelihood problems. At each iteration one updates the current estimate of the true likelihood maximiser, $\hat{\theta}_k$, to a better estimate, $\hat{\theta}_{k+1}$. Provided that $Q(\hat{\theta}_{k+1},\hat{\theta}_k) > (\hat{\theta}_k,\hat{\theta}_k)$ then the algorithm guarantees that the likelihood associated with the new estimate will be strictly

greater than that of the old one. Thus, iterating the algorithm produces a sequence of estimates $\{\hat{\theta}_k\}$ associated with a monotonically increasing sequence of likelihoods $\{L(\hat{\theta}_k)\}$. The key idea behind the approach is to simplify each iteration by introducing an extra degree of freedom. The EM algorithm allows the user to select a set of unobserved, yet desirable, missing data which, in addition to the actual observations, constitute the so-called complete data set. Normally the user would choose the missing data so that maximising the likelihood function associated with the complete data set is easy. Often this problem is then solvable in closed-form. At each iteration, one calculates the maximiser of the projection of this complete data likelihood onto the space of actual observations in directions informed by the estimate from the previous iteration.

It turns out that the role of the missing data is more crucial to the success of the algorithm than at first glance. Indeed, as revealed in Section (4.1.2), when the algorithm converges to some particular estimate then the missing data plays an important role in determining the speed of convergence of the algorithm to that estimate. Furthermore, the rate of convergence is a function of the eigenvalues of the expected augmented information matrix. Finally, note that if the problem of maximising the log-likelihood function for the observed data is difficult and that of maximising $Q(\theta, \hat{\theta}_k)$ simple, then via the fundamental equation (4.13), it follows that maximising $log\, k(z \mid y, \theta)$ must also be difficult. The beauty of the approach is that the EM algorithm never requires the explicit computation of $log\, k(z \mid y, \theta)$.

Appendix A

The Proof of Lemma 15

Proof. From (4.13) we have

$$\frac{\partial L(\theta)}{\partial \theta} \mid_{\theta=\hat{\theta}_{k+1}} = \frac{\partial Q(\theta, \hat{\theta}_k)}{\partial \theta} \mid_{\theta=\hat{\theta}_{k+1}} - \frac{\partial \nu(\theta, \hat{\theta}_k)}{\partial \theta} \mid_{\theta=\hat{\theta}_{k+1}} . \tag{4.24}$$

where $\nu(\theta, \hat{\theta}_k) = log\, k(z \mid y, \theta)$.

The first term on the right-hand side of (4.24) is zero by assumption (3), while the second term is zero in the limit as $k \to \infty$, and hence (4.18).

Similarly, $\frac{\partial^2}{\partial\theta\partial\theta^T} Q(\theta, \theta^*) \mid_{\theta=\theta^*}$ is negative definite, since it is the limit of the sequence

$$\frac{\partial^2}{\partial\theta\partial\theta^T} Q(\theta, \hat{\theta}_k) \mid_{\theta=\hat{\theta}_{k+1}},$$

each of whose terms has all eigenvalues bounded away from zero.
Finally, we turn to the problem of establishing (4.19). Expanding $\frac{\partial^2}{\partial\theta\partial\theta^T} Q(\theta, \theta_1) \mid_{\theta=\theta_1}$ in a Taylor series about the point (θ^*, θ^*) yields

$$\begin{aligned} \frac{\partial Q(\theta, \theta_1)}{\partial\theta} \mid_{\theta=\theta_2} \;=\; & \frac{\partial Q(\theta, \theta^*)}{\partial\theta} \mid_{\theta=\theta^*} + \Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big](\theta_2 - \theta^*) \\ & + \Big[\frac{\partial^2 Q(\theta, \gamma)}{\partial\theta\partial\gamma^T} \mid_{\theta=\theta^*, \gamma=\theta^*}\Big](\theta_1 - \theta^*) + \ldots \end{aligned} \tag{4.25}$$

Substituting $\theta_1 = \hat{\theta}_k$ and $\theta_2 = \hat{\theta}_{k+1}$ into equation (4.25) we obtain

$$0 = \Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big](\hat{\theta}_{k+1} - \theta^*) + \Big[\frac{\partial^2 Q(\theta, \gamma)}{\partial\theta\partial\gamma^T} \mid_{\theta=\theta^*, \gamma=\theta^*}\Big](\hat{\theta}_k - \theta^*) + \ldots$$

Since $\hat{\theta}_{k+1} = M(\hat{\theta}_k)$ and $\theta^* = M(\theta^*)$ we obtain in the limit

$$0 = \Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big]\Big[\frac{\partial\, M(\theta)}{\partial\theta} \mid_{\theta=\theta^*}\Big] + \frac{\partial^2 Q(\theta, \gamma)}{\partial\theta\partial\gamma^T} \mid_{\theta=\theta^*, \gamma=\theta^*},$$

or

$$\frac{\partial\, M(\theta)}{\partial\theta} \mid_{\theta=\theta^*} = -\Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big]^{-1}\Big[\frac{\partial^2 Q(\theta, \gamma)}{\partial\theta\partial\gamma^T} \mid_{\theta=\theta^*, \gamma=\theta^*}\Big]. \tag{4.26}$$

Now, employing (4.13) yields

$$\frac{\partial\, M(\theta)}{\partial\theta} \mid_{\theta=\theta^*} = -\Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big]^{-1}\Big[\frac{\partial^2 \nu(\theta, \gamma)}{\partial\theta\partial\gamma^T} \mid_{\theta=\theta^*, \gamma=\theta^*}\Big], \tag{4.27}$$

Finally

$$\frac{\partial\, M(\theta)}{\partial\theta} \mid_{\theta=\theta^*} = \Big[\frac{\partial^2 Q(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big]^{-1}\Big[\frac{\partial^2 \nu(\theta, \theta^*)}{\partial\theta\partial\theta^T} \mid_{\theta=\theta^*}\Big], \tag{4.28}$$

Appendix B

Taylor's Theorem, Residual Errors and the Logarithm

This section looks at results related to the well-known Taylor s Theorem. First, we shall state this theorem.

Lemma 16 ***Taylor's Theorem*** *If a function f has n continuous derivatives on the interval* $[a, b]$ *and its* $(n+1) - st$ *derivative exists on the interval* (a, b)*, then for* $x_0 \in [a, b]$ *and* $x \in [a, b]$*,*

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f^1(x_0) + \frac{(x - x_0)}{2}f^{(2)}(x_0) + \ldots \frac{(x - x_0)}{n!}f^{(n)}(x_0) \\ &\quad + \frac{1}{n!}\int_{x_0}^{x}(x - \epsilon)^n f^{(n+1)}(\epsilon)d\epsilon. \end{aligned} \tag{4.29}$$

Now, a function f, satisfying the conditions of Taylor's Theorem, may be written as

$$f(x) = T_n(x) + R_n(x),$$

where $T_n(x)$ is the Taylor series and $R_n(x)$ the remainder term. These are implicitly defined by equation (4.33).

Unfortunately, the expression the residual error,

$$R_n(x) = \frac{1}{n!}\int_{x_0}^{x}(x - \epsilon)^n f^{n+1}(\epsilon)d\epsilon,$$

is not terribly convenient since it contains an integration operator. It is possible to derive a more useful expression with which to quantify the residual error.

The derivation of this expression begins with the introduction of an auxiliary function $F(t)$, which is defined as follows

$$\begin{aligned} F(t) &= f(x) - f(t) - f^1(t)(x - t) - \frac{f^{(2)}(t)}{2!}(x - t)^2 - \ldots \\ &\quad - \frac{f^{(n)}(t)}{n!}(x - t)^n - K(x - t)^{n+1}. \end{aligned} \tag{4.30}$$

Here the constant K is chosen so that $F(x_0) = 0$.

Now, since $F(x) = 0$, by Rolle's Theorem there must be a point $\lambda \in (x_0, x)$ such that

$$F^1(\lambda) = 0. \tag{4.31}$$

Since $F^1(\lambda)$ is given by

$$\begin{aligned} F^1(t) &= -f^1(t) + \big[f^{(1)}(t) - f^{(2)}(t)(x-t)\big] + \big[f^{(2)}(t)(x-t) - \frac{1}{2!}f^{(3)}(t)(x-t)^2\big] \\ &\quad + \big[\frac{1}{2!}f^{(3)}(t)(x-t)^2 - \frac{1}{3!}f^{(4)}(t)(x-t)^3\big] + \dots \\ &\quad + \big[\frac{1}{n-1!}f^{(n)}(t)(x-t)^{n-1} - \frac{1}{n!}f^{(n+1)}(t)(x-t)^n\big] + (n+1)K(x-t)^n \\ &= (n+1)K(x-t)^n - \frac{1}{n!}f^{(n+1)}(x-t)^n, \end{aligned} \tag{4.32}$$

equation (4.31) implies that

$$(n+1)K(x-\lambda)^n - \frac{f^{n+1}(\lambda)}{n!}(x-\lambda)^n = 0,$$

and thus

$$K = \frac{f^{n+1}(\lambda)}{n+1!}.$$

Finally, substituting this value for K into (4.30) and setting $t = x_0$ in (4.30) provides

$$\begin{aligned} 0 &= f(x) - f(x_0) + (x-x_0)f^1(x_0)\frac{(x-x_0)}{2}f^{(2)}(x_0) - \dots \\ &\quad - \frac{(x-x_0)}{n!}f^{(n)}(x_0) - \frac{(x-x_0)}{(n+1)!}f^{(n+1)}(\lambda), \end{aligned} \tag{4.33}$$

which is Taylor's formula (4.33).

The result of this derivation is now presented as a corollary.

Corollary 17 *If a function f has n continuous derivatives on the interval* $[a,b]$ *and its* $(n+1)^{st}$ *derivative exists on the interval* (a,b)*, then for* $x_0 \in [a,b]$ *and* $x \in [a,b]$*,*

$$\begin{aligned} f(x) &= f(x_0) + (x-x_0)f^1(x_0)\frac{(x-x_0)}{2}f^{(2)}(x_0) + \dots - \frac{(x-x_0)}{n!}f^{(n)}(x_0) \\ &\quad + \frac{(x-x_0}{(n+1)!}f^{(n+1)}(\lambda), \end{aligned} \tag{4.34}$$

for some $\lambda \in [x_0, x]$.

Chapter 5

Bayesian Adaptive Filtering

After reviewing the generalities in the previous chapters about sate estimation and its parameters estimation, we tackle in this chapter the problem of adaptive filtering in non-stationary environments. We first begin by an introductory state of the art review which includes the existing algorithms on adaptive filtering (LMS and LRS). These algorithms experience performances limitation in terms of tracking and convergence in non-stationary environments. This motivates our work to propose more efficient technics for such Bayesian technics. Our proposed methods take into consideration a priori information about the system variations such as the PDP, Doppler bandwidth, ...etc. We thus propose two different approaches, the first one is based on Wiener Filtering (WF) while the other one is based on Kalman Filtering (KF). In this chapter we develop the first approach and the second one will be considered in the next chapter. The proposed algorithms can be applied in many situations and as an example we consider in this chapter its application for system identifications in particular mobile radio channel for the importance of this medium in wireless communications. We provide numerical results that show the proposed algorithm advantage compared to existing algorithms in terms of Excess Mean Squared Error (EMSE) for different PDPs and Doppler shifts.

5.1 State of Art

Since the introduction of the LMS algorithm by Widrow and Hopf in the 1960's, most of the further work in adaptive filtering has focused on improving the initial convergence. The Recursive Least-Squares (RLS) algorithm was also developed in the 1960's and provided an alternative algorithm for adaptive system identification. The RLS algorithm is recursive and not iterative as the LMS algorithm, solving a LS cost function exactly at each update. As a result it converges very fast since it provides an unbiased solution once the LS problem gets overdetermined. This deterministic aspect adds up to the observation that the RLS convergence is insensitive to the input signal correlation structure (approximately, since there is some dependence on the initialization). The RLS algorithm, though providing computational savings w.r.t. the plain solving of LS problems at each sampling period, is quite a bit more expensive than the LMS algorithm. This motivated on the one hand the development of fast RLS algorithms, and on the other hand the development of an intermediate category of algorithms, all less sensitive than LMS to the input correlation structure, including frequency or other transform domain LMS algorithms, prewhitened LMS versions, Fast Newton Transversal Filters and (Fast) Affine Projection Algorithms.

At the outset, all these algorithms were developed to converge to an unknown optimal filter. When this optimal filter is actual time-varying, these algorithms need to be made adaptive. The RLS algorithms are made adaptive by the introduction of a weighting function/window. The weighted LS cost function can be viewed as the output of a filter with the instantaneous squared filtering error sequence as input. The filter should be such that its input-output relationship is simple and recursive. The LS cost function uses a discrete-time integrator as filter, which can be easily modified into a first-order recursive filter for the exponentially weighted RLS algorithm. The sliding window RLS algorithm uses a moving average filter that can also be expressed recursively. All other adaptive filtering algorithms are made adaptive by the introduction of a scalar stepsize. In fact, the time-varying stepsize sequence of stochastic gradient algorithms [10] is made time-invariant/constant to avoid convergence and permit tracking of time-varying optimal filter settings. The tracking characteristics of the LMS and RLS algorithms got analyzed only in the 1970's and 1980's, 10 to 20 years after the introduction of the algorithms,

in [22] for LMS and [21] for RLS. A further inspection of these tracking characteristics revealed the surprising result that in certain cases the LMS algorithm may provide better tracking than the RLS algorithm (each with optimized stepsize or forgetting factor), see [23] for deterministic and e.g. [10] for random parameter variations. With hindsight, this is not at all surprising since LMS and RLS are just two suboptimal approaches to tracking time-varying parameters. Whereas initial convergence is about the fast reduction of the mean parameter error vector, tracking is about the optimal compromise between MSE due to estimation noise and tracking/lag noise. Some general references on the tracking behavior of adaptive filtering algorithms are [1, 4, 3], [6, 16, 19] and [17, 18, 14].

The RLS algorithm got introduced after the Kalman filter (KF) was invented, though the RLS algorithm is a special case of the KF for the following state-space model [13]

$$H_k = H_{k-1} \tag{5.1}$$

$$x_k = Y_k^T H_k + v_k \ . \tag{5.2}$$

The KF formulation requires immediately a parametric form of the optimal filter, usually a FIR filter is assumed with impulse response of N coefficients contained in the vector H_k. The measurement equation (5.2) expresses that the desired-response signal x_k is the sum of the output $\sum_{i=0}^{N-1} H_{k,i} y_{k-i}$ of the optimal FIR filter H_k with *input* y_k plus an independent *measurement noise* v_k. In KF terminology, x_k would be the *measurement* and H_k the *state*.

Wiener filtering (WF'ing) is about estimating one random signal from another, let's say estimating the signal x_k on the basis of the signal y_k. In the system identification set-up of adaptive filtering, which is reflected in (5.1)-(5.2), the relation between these two signals is that x_k is assumed to be output of an unknown system/plant with y_k as input plus independent measurement noise v_k. In this case, the optimal Wiener (LMMSE) filter is clearly $R_{xY} R_{YY}^{-1} = H^T$, which is FIR if the system to be identified is FIR. The WF is based on the joint statistical description of the random signals x_k and y_k. and is a deterministic quantity. The WF solution is not influenced by the color of the noise v_k.

KF'ing is in principle a special case of the signal-in-noise case of WF'ing. In the signal-in-noise case, the measurement signal is the sum of the signal to be estimated plus noise. For the KF, the signal to be estimated satisfies furthermore a state-space model.

The adaptive filtering/RLS application of KF'ing though deviates significantly from this spirit. In RLS, the quantity (state) estimated is the set of WF coefficients H_k instead of its output, the filter input y_k is considered deterministic (the estimation is given y_k) and hence the filter estimate would be random if y_k would be considered random. Indeed, the Kalman Filter provides an estimate $\hat{H}_k$ of the WF H_k. Since this KF application is now an instance of parameter estimation, the parameter estimation quality depends on e.g. the color of the noise v_k.

The KF'ing framework can be straightforwardly extended to incorporate time-varying optimal parameters. The simplest way is probably through the following stationary AR(1) model state equation for the optimal filter variation [13]

$$H_k \;=\; A\, H_{k-1} + W_k \tag{5.3}$$

replacing (5.1), where $\mathrm{E}\; W_k W_i^H \;=\; Q\,\delta_{ik},\;\; \mathrm{E}\; W_k v_i^H \;=\; 0$ (noises assumed circular in complex case). This formulation lead to the widely accepted point of view that the KF would be the optimal adaptive filter. This is indeed true for the system identification configuration with (5.3)-(5.2) as assumed correct model and A, Q and r in $\mathrm{E}\; v_k v_i^H = r\,\delta_{ik}$ assumed known. We may note that in this model, WF'ing provides the time-varying optimal filter $H_k^T = R_{x_k Y_k} R_{Y_k Y_k}^{-1}$ and the Kalman Filter estimates it in a Bayesian (LMMSE) sense.

The problem with the KF viewpoint is that the model parameters, if at all the model is correct, are unknown and need to be estimated also from the same data. Those parameters can be inferred from the joint signal statistics, just like the Wiener Filter itself. However, in the KF, the input signal y_k is considered deterministic which makes the state space model (5.3)-(5.2) linear but time-varying. These complications lead to approximate approaches such as exponentially weighted RLS, which can be shown [7] to correspond to the KF for certain artificial choices of A and Q in (5.3). The main issue in most applications is the so-called generalization property of statistical learning: what counts is the adaptive filter performance not for the given input signal realization, but when applied to other signal data, hence for the given signal statistics. The generalization capacity may be hampered by sticking too closely to one model's details when the model is approximate. Another issue is that the Kalman Filter approach for tracking time-varying optimal filters only applies in the system identification configuration in which the filter's non-stationarity

arises in the cross-correlations between input and desired-response signals, regardless of the statistics ((non)stationarity) of the input. Communications applications of the system identification configuration are channel estimation and echo cancelation. In all other configurations of adaptive filtering: prediction, deconvolution/equalization and interference cancelation, the statistics of the optimal filter may be strongly intertwined with the statistics of the input signal. In linear prediction for instance, the desired-response and input signals are the same. One rarely sees the linear prediction problem addressed as a ML estimation of or Kalman Filtering on the parameters of an AR model, because any AR model order is likely to lead to an approximation error. Adaptive prediction is in fact a joint operation of approximation (e.g. through model order selection) and estimation. In equalization, even if the channel variation could be modeled as an AR(1) model as in (5.3), the optimal equalizer setting is a nonlinear function of the channel. Given all these considerations, the best practical approach is probably to specify a motivated solution structure of acceptable complexity and optimize the parameters within that structure (as is done in linear prediction) (approximation/estimation compromise). The problem considered here has of course been addressed previously and we now discuss some of this existing work.

5.1.1 Tracking Bandwidth Adjustment

Most of the work on adapting tracking capability has focused on adapting one tracking parameter. In RLS, it doesn't cost any computational complexity to make the forgetting factor (FF) time-varying. Modifications to fast RLS algorithms to allow a time-varying FF, as well as algorithms to adjust this FF on the basis of correlation matching have been pursued in [82]. The equivalent development for LMS algorithms concerns Variable StepSize (VSS) algorithms. Important developments were presented in [37], [39], [65], [64], [66] [38] and [42]. Most of the VSS algorithms use the steepest-descent strategy and the instantaneous squared error cost function of the LMS algorithm to adjust the additional parameter, which is the stepsize. A related but different approach consists in running various adaptive filters with different time constants and selecting or combining their outputs, similarly to what is done in model order selection, see [71], [70], [68] and [69].

A further refinement is to allow different tracking bandwidths for different filter components as is done in [40] with a VSS per filter coefficient and in [81] where the tracking capacity increases with frequency for the various frequency domain components of the filter. The work in [40] essentially shows that a "diagonal" state-space model (5.3) may allow a simplification of the KF to a LMS algorithm with a VSS per tap, but no attempt is made to automatically adjust the resulting stepsizes.

5.1.2 Power Delay Profile

Besides the statistical modeling of the parameter variation, another important ingredient in Bayesian adaptive filtering is the incorporation of prior knowledge on the coefficient sizes. Indeed, when tracking time-varying filters, it becomes possible to learn the variances of the filter coefficients. This aspect has been exploited for a while in a rudimentary, binary form for sparse filters: filter coefficients are either adapted or deemed to small and kept zero (for each filter coefficient, the stepsize is either 0 or a constant). More recently, a smoother evolution of the stepsize has been introduced, leading to the Proportionate LMS (PLMS) algorithm, motivated e.g. by acoustic echo cancelation in which the adaptive filter has many coefficients, but their value tapers off, see [78],[79]. Similar prior information is starting to be taken into account for (LMMSE) channel estimation in wireless communications [84], where the evolution of the channel coefficient variances along the impulse response is called the power delay profile, important developments were presented in [27, 32, 27].

5.1.3 Full Bayesian Approach

In a full Bayesian approach, the whole matricial spectrum $S_H(z) = S_{HH}(z)$ of H_k counts: not only the parameter variation speed/bandwidth but the whole spectral shape counts, not only the spectral shape but also the power delay profile counts, and in principle also the cross spectra between coefficients need to be accounted for.

The KF [13] allows to do all this in the system identification set-up, but ignores the es-

timation of $S_H(z)$. In [77], a point of view close to the one of this is developed. However, they require the knowledge of the (multivariate IIR) matricial spectrum of the (standard) adaptive filter gradient (this could be estimated from the observations of the gradient) and knowledge of the (multivariate IIR) matricial spectrum of the stationary filter parameter vector. This last requirement is quite unrealistic. Furthermore, the design steps suggested may be quite sensitive to estimation errors to some quantities that get estimated.

5.2 System Identification

Consider now the prototype adaptive filtering set-up, which is the system identification set-up, in which the desired response signal y_k is modeled as the output of the optimal filter, which can be time-varying, plus independent (white) noise. The adaptive system identification Fig. 6.1 is designed for determining a (typically linear FIR) model of the transfer function for an unknown, time-varying digital or analog system.

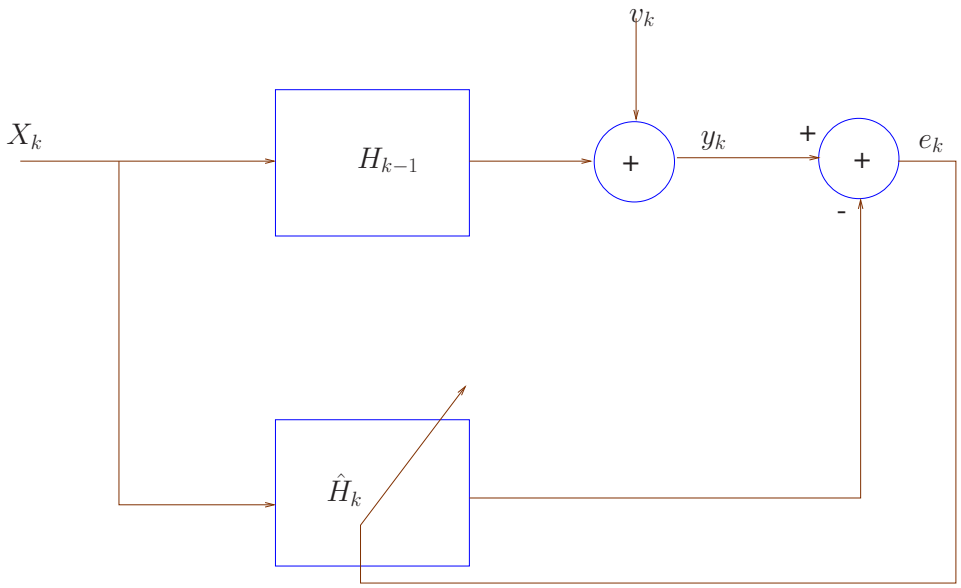

Figure 5.1: System identification block diagram

For the performance analysis, we will assume that the adaptive filter structure is that of an N-point FIR filter, and the input signal X_k is obtained as a vector formed by the

most recent N samples of the input sequence x_k, i.e.,

$$X_k = [x_k,\ x_{k-1}, \ldots, x_{k-N+1}]^T. \tag{5.4}$$

Let H_k denote the optimal coefficient vector (in the minimum mean-squared estimation error sense (MMSE)) for estimating the desired response signal y_k using X_k. We will assume that $H(k)$ is time varying, and that the time variations are caused by a random disturbance of the optimal coefficient process.
In order to make the analysis tractable, we will make use of the following assumptions and approximations:

- X_k, y_k are jointly Gaussian and zero-mean random processes. X_k is a stationary process. Moreover, $\{X_k,\ y_k\}$ is uncorrelated with $\{X_n,\ y_k\}$ if $n \neq k$. This is the commonly employed independence assumption and is seldom true in practice. However, analysis reliable design rules in the past.

- The autocorrelation matrix R_{XX} of the input vector X_k is a diagonal matrix and is given by:

  $$R_{XX} = \sigma_x^2 I. \tag{5.5}$$

  While this is a fairly restrictive assumption,it considerably simplifies the analysis. Furthermore, the white data model is valid representation in many practical systems such as digital data transmission systems and analog systems that are sampled at Nyquist rate and adapted using discrete-time algorithms.

5.3 Modeling of Standard Adaptive Filtering Behavior

The adaptive filter is $\hat{H}_k$ and the a prior error $e_k = y_k - X_k^H \hat{H}_{k-1}$. Consider the (complex) LMS algorithm first

$$\begin{aligned} H_k^{lms} &= H_{k-1}^{lms} + \mu X_k^H e_k \\ &= (I - \mu X_k^H X_k) H_{k-1}^{lms} + \mu X_k^H v_k + \mu X_k^H X_k \\ &= (I - \mu X_k^H X_k) H_{k-1}^{lms} + \mu X_k^H v_k + \mu R H_k \mu (R - X_k^H X_k)(H_{k-1}^{lms} - H_k) \end{aligned} \tag{5.6}$$

Then, assuming the adaptation speed is not too fast, we get approximately

$$H_k^{lms} = [I - (I - \mu R)q^{-1}]^{-1}\mu R(H_k + R^{-1}X_k^H v_k) \tag{5.7}$$

whereas the RLS filter update is of the form

$$\begin{aligned} H_k^{rls} &= H_{k-1}^{rls} + \hat{R}^{-1}X_k^H e_k \\ &= \frac{1-\lambda}{1-\lambda q^{-1}}(H_k + R^{-1}X_k^H v_k) \end{aligned} \tag{5.8}$$

In general

$$\begin{aligned} \hat{H}_k &= F_{lms,rls}(q)(H_k + R^{-1}X_k^H v_k) \\ &= F(q)G_k \end{aligned} \tag{5.9}$$

Gradient $G_k = R^{-1}X_k^H y_k$ in fact! then we can estimate S_{GG} assume RLS or LMS with white, where $q^{-1}H_k = H_{k-1}$. Using averaging analysis at low adaptation speed, these results for the sysid-up hold approximately also for the other adaptive filtering applications. Note that $(H_k + R^{-1}X_k^H v_k)$ is closely related to $G_k = R^{-1}X_k^H y_k$, which is a mixed quantity in that it is averaged in the input covariance but instantaneous in the input-desired-response correlation.

5.4 Bayesian Adaptive Filtering (BAF)

In this chapter we focus on stationary time-varying parameters, we neglect transient phenomena, and we consider the stationary steady-state regime. Hence we formulate the parameter tracking problem as a Wiener filtering problem.

5.4.1 Wiener solution

We shall introduce, mostly for the purpose of analysis, a somewhat idealized Bayesian solution which is based on the assumption that R can be estimated well. This solution

will be based on LMMSE estimation (WF'ing) of H_k from the gradient:

$$G_k = R^{-1}X_k^H y_k \;=\; H_k + \underbrace{R^{-1}X_k^H v_k}_{\tilde{G}_k} + (R^{-1}X^H X - I)\, H_k \tag{5.10}$$

where for slow parameter variations, the last term can be neglected since it is the product of low-pass noise H_k with high-pass noise $R^{-1}X_k^H X_k - I$. The optimal BAF would be to apply the kF to (11), $G_k = H_k + \tilde{G}$, which can be considered as a measurement equation for the state H_k. In steady-state, the Kf converge to the WF

$$H_k = F(q)G_k \tag{5.11}$$

where in the non-causal case

$$F(q) = I - S_{\tilde{G}\tilde{G}}(q)S_{GG}^{-1}(q) \tag{5.12}$$

Neglecting the last term in (5.10) and assuming that v_k is white noise (hence $\tilde{G}_k$), we have $S_{\tilde{G}\tilde{G}}(q) = \sigma_v^2 R^{-1}$. Hence the non-causal WF is fairly straightforward to find since S_{GG} can be estimated simply from the observations of G_k, though σ_v^2 is somewhat trickier to derive from the observed MSE.

For the causal case, consider $N_k = P(q)G_k$ where $P(q)$ is the (length) (monic) multivariate prediction error filter for the vector signal G_k and N_k is resulting white prediction error with covariance matrix R_{NN}. then the causal WF is

$$F(q) = I - S_{\tilde{G}\tilde{G}}(q)R_{NN}^{-1}P(q) \tag{5.13}$$

5.5 Application: Mobile Radio Channel

In mobile radio communication, the transmission between a transmitter (T_x) and a mobile receiver (R_x) takes place via many paths. A direct wave reaches the receiver if a line-of-sight (LOS) path exists. Other waves are scattered, reflected, or diffracted at natural or man-made obstacles. Hence, these waves are characterized by **attenuation, time delay,**

and angle of- arrival.

The multipath propagation influences the resulting signal in several ways such as time dispersion, fading, and Doppler shift. In this section, these effects are explained, and the channel model for the simulation is introduced.

Time Dispersion The signature displaying the received signal energy versus time delay is called the PDP. The PDP describes the time dispersion due to multipath propagation. Shape and length of the PDP are affected by the nature of the environment, particularly by the size and density of buildings or other obstacles.

In general, it can be assumed that short propagation paths occur more often than longer propagation paths, yielding a higher contribution to the entire received signal energy. According to measurements , this effect can be described by an approximately negative exponential variation of the received signal. Very often, the influence of a certain obstacle can be discerned directly in the PDP shape. Irregularities of the PDP described by peaks or echo groups occur [36]. The different terrain types are usually separated into the groups urban, suburban, and rural, corresponding to their building density [30]. These qualitative descriptions are not precise and are open to different interpretation. Due to the strong variation of realistic channels, a clear-cut separation with respect to the geographic situations is impossible [35]. Therefore, this chapter focuses on the influence of the details of a more general PDP and Doppler shift instead of investigating the representatives of a special area.

Fading All incoming waves combine vectorially at the receiving antenna. At a moving receiver, the phases and amplitudes of the different multipath signals will change rapidly due to local scatterers. The resulting spatial field pattern varies from point to point. The probability density function (pdf) of the signal envelope variation can be approximated by a Rayleigh distribution, if all contributing signals have about the same magnitude and an equally distributed random phase.

When moving into shadows of buildings for example, the total path loss changes due

to large local shadowing effects like buildings or trees. As these effects occur slowly, compared to the fast fading described above, these fluctuations of the mean level of the received signals are called slow fading. Measurements show that this slow fading can be approximated by a log-normal distribution. In general, the parameters of the log-normal pdf depend on the environment type. Hence, the standard deviation varies in the range of $2 - 6dB$ [31].

Doppler shift Whenever relative motion exists between T_x and R_x , an apparent shift in the frequency of the received signal occurs due to Doppler shift. The Doppler shift is different for every incoming wave as it is related to the velocity component of the vehicle in the incoming direction. This yields a complete Doppler spectrum. Otherwise, assumptions about the angle-of-arrival or corresponding reference Doppler spectra should have been made. Consequently, the results gained in this chapter are valid for the static channel.

Channel Model We consider the uniform dynamics plus power delay profile like structured model for the optimal Doppler spectrum: $S_{HH}(e^{j2\pi f}) = S_{hh}(e^{j2\pi f})\ D$ where $S_{hh}(e^{j2\pi f})$ is scalar, f represent the Doppler shift and the matrix D is arbitrary if it were diagonal (decorrelation filter coefficients) the diagonal would represent the power delay profile(PDP) of the optimal filter
To simplify, we suppose, also, that the scalar spectrum S_{hh} is a flat low-pass spectrum; i.e.

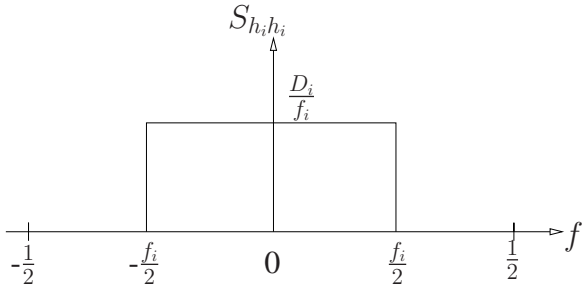

Figure 5.2: Doppler spectrum representation

5.6 Performance Analysis

In this section we will comparing the performance between the SAF and BAF in terms of the resulting EMSE.

The EMSE is defined by

$$\begin{aligned} EMSE &= E\left[e_k^2\right] - MMSE \\ &= E\left[X_k^H \widetilde{H}_k^* \widetilde{H}_k^T X_k\right] \end{aligned} \tag{5.14}$$

where

$$\begin{aligned} \tilde{H}_k &= H_k - H_k \\ &= (I - F(e^{-j2\pi f}))\ H_k\ -\ R^{-1}\ F(e^{-j2\pi f}) X_k^* v_k \\ &= (I - F(e^{-j2\pi f}))\ H_k\ -\ F(e^{-j2\pi f}) \tilde{G}_k \end{aligned}$$

If we make the assumption that the system variation is a zero-mean, wide-sense stationary process with a cross-spectral density matrix $S_{HH}(e^{-2j\pi f})$, and if we suppose that these variations are independent from the input signal, the Excess MSE can be expressed in the following form:

$$\begin{aligned} EMSE &= tr\{E(\widetilde{H}_k \widetilde{H}_k^H R)\} \\ &= tr\{R \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{\tilde{H}\tilde{H}^H}(e^{j2\pi f}) df\} \end{aligned} \tag{5.15}$$

and becomes:

$$\begin{aligned} EMSE &= trR\{\int_{-\frac{1}{2}}^{\frac{1}{2}} F(e^{j2\pi f}) R_{\tilde{G}\tilde{G}}\ F^H(e^{j2\pi f}) df\} \\ &\quad + trR\{\int_{-\frac{1}{2}}^{\frac{1}{2}} (I - F(e^{j2\pi f}))\ \times S_{HH}\ (e^{-j2\pi f})(I - F^H e^{j2\pi f})\} df \end{aligned}$$

Remark that the EMSE can be broken up into two terms:

- $E_{noise} = trR\{\int_{-\frac{1}{2}}^{\frac{1}{2}} F(e^{j2\pi f}) R_{\tilde{G}\tilde{G}}\ F^H(e^{j2\pi f}) df\}$ characterizing the noise contribution; and can be interpreted as the estimation accuracy under stationary conditions
- $E_{lag} = trR\{\int_{-\frac{1}{2}}^{\frac{1}{2}} (I - F(e^{j2\pi f}))\ S_{HH}(e^{j2\pi f}))(I - F^H(e^{j2\pi f}))\} df$ representing the estimation error resulting from the system variations (Lag noise)

in the **RLS case**

$$F(z) = \frac{1-\lambda}{1-\lambda z^{-1}} I, \tag{5.16}$$

in the **LMS case**

$$F(z) = \frac{\mu\sigma_x^2}{1-(1-\mu\sigma_x^2)z^{-1}} I, \tag{5.17}$$

in the **non-causal case**

$$F(z) = I - S_{\tilde{G}\tilde{G}}(z) S_{GG}^{-1}(z) \tag{5.18}$$

in the **causal case**

$$F(z) = I - S_{\tilde{G}\tilde{G}}(z) R_{NN}^{-1} P(z) \tag{5.19}$$

We deduce, thus, the EMSE expressions for the different are given by (5.20) for different windows:

$$EMSE^{RLS} = N\sigma_v^2 \frac{1-\lambda}{1+\lambda} + 2\sigma_x^2 tr(D)\big(\lambda f_i - \frac{\lambda}{\pi}\frac{1-\lambda}{1+\lambda}\arctan(\frac{1+\lambda}{1-\lambda}\tan(\pi f_o))\big) \tag{5.20}$$

$$EMSE^{LMS} = N\sigma_v^2 \frac{\mu\sigma_x^2}{2-\mu\sigma_x^2} + 2\sigma_x^2(1-\mu\sigma_x^2) tr(D)\big(f_i - \frac{\mu\sigma_x^2}{\pi(2-\mu\sigma_x^2)}(\arctan\frac{\mu\sigma_x^2}{2-\mu\sigma_x^2}\tan(\pi f_i))\big) \tag{5.21}$$

$$EMSE_{ncc} = \sum_{i=1}^{N} \frac{1}{j2\pi} \oint \frac{dz}{z} (\frac{1}{\sigma_v^2} + \frac{1}{\sigma_x^2} D_i S_{hh}(z))^{-1} = \sigma_v^2 2 f_i \sum_{i=1}^{N} \frac{1}{1 + \frac{\sigma_v^2}{\sigma_x^2} D_{ii} \frac{1}{2f_i}} \tag{5.22}$$

$$EMSE_{cc} = \sigma_v^2 2 f_i \sum_{i=1}^{N} [1 - (1 + \frac{\sigma_x^2 D_i}{\sigma_v^2 f_i})^{-f_i}] \tag{5.23}$$

Proof. see the appendix.

5.7 Numerical Results

In this section the behavior of BAF and standard adaptive filters is compared for non-stationary environments in a system identification setup. We consider the particular scenario of **Wireless Radio Channel**. In all simulations presented here, the desired signal y_k is corrupted by zero mean, (iid) Gaussian noise of σ_v^2 variance.
We compare the minimum EMSE achieved by each variable (with optimized parameters $\lambda\,;\mu$) and by a BAF.
Fig. (5.3) plots the minimum EMSE curves as a function of the power delay profile D_i, for a fixed small Doppler bandwidth ($f_i = 0.001$). We see that the Baysian Adaptive Filtering (BAF) given with a causal and non-causal Wiener filter performs better and the optimal RLS and LMS have bad performance for a small Doppler bandwidth.
Fig. (5.4) plots the EMSE as a function of the power delay profile D_i, for a fixed Doppler bandwidth ($f_i = 0.1$). We can notice that for optimal step-size in the LMS and optimal $\lambda = 0.97$ in the RLS, the two algorithms show good performances. However, for a large step size, the LMS algorithm does not track well, while the BAF outperforms the SAF for different Doppler shifts.
Fig. (5.5) plots the EMSE as a function of the power delay profile D_i and Doppler bandwidth f_i. This curve shows that the BAF performs even better than SAF.

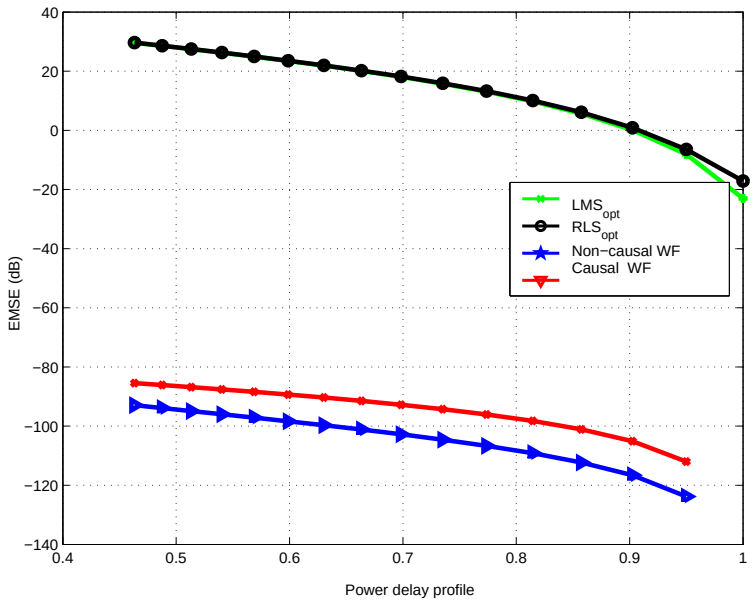

Figure 5.3: Comparative tracking performance results between BAF and Standard AF using EMSE for different value of power delay profile at Doppler bandwidth ($f_i = f_0 = 0.001$)

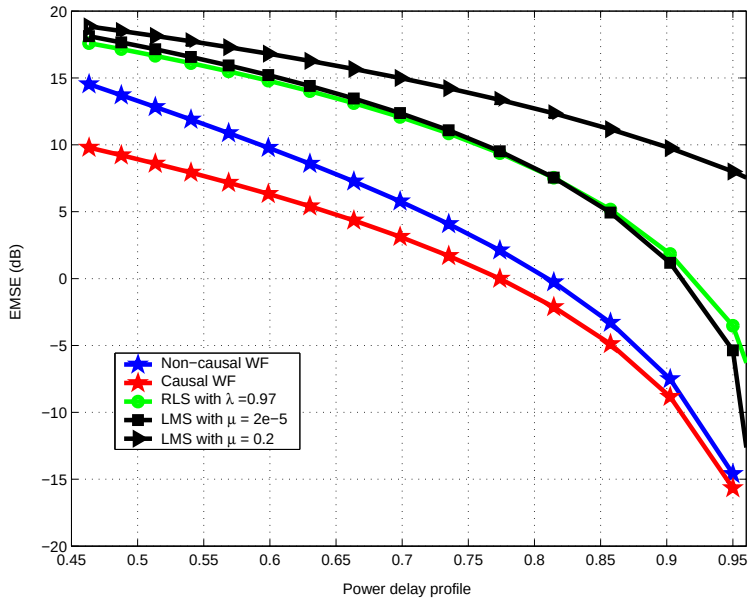

Figure 5.4: Comparative tracking performance results between BAF and Standard AF using EMSE for different value of power delay profile at Doppler bandwidth ($f_i = f_0 = 0.1$)

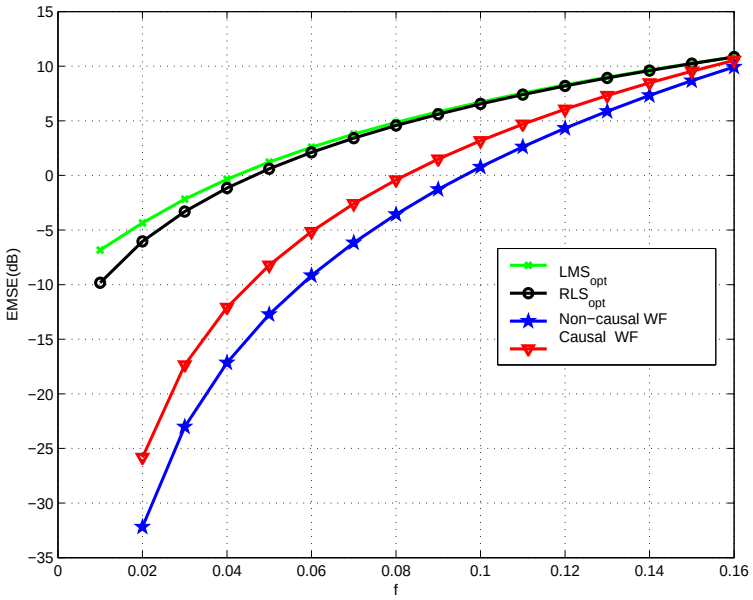

Figure 5.5: Comparative tracking performance results between BAF and Standard AF using EMSE for different value of power delay profile and a different value Doppler bandwidth (f_i)

5.8 Concluding remarks

In this chapter we proposed a bayesian method based on WF tacking into account priori information in order to improve the tracking and convergence performance. We also consider the application of our proposed method in the case of mobile radio communications where the priori information consist on PDP and Doppler shift. The numerical results show the impact of the priori information on the performance of the propped technique. In the following, we summarize the main characteristics of our technique

- For a small Doppler Bandwidth (DB) the optimal RLS and LMS track poorly for different values of PDP.
- For a large Doppler Bandwidth, the LMS with a small step-size and RLS with optimal λ track well but do not outperforme the BAF. Also for a large step-size the LMS does not track properly.
- For the different aprior information (DB and PDP), the standard adaptive filtering

(LMS and RLS) are not bad, but the BAF still show the best performance.

In the following chapter we propose and other Bayesian technique based on KF.

APPENDIX

EMSE for a causal WF

The z-transform of the causal WF window is given by:

$$F(q) = I - S_{\tilde{G}\tilde{G}}(q) R_{NN}^{-1} P(q) \tag{5.24}$$

where $N_k = P(q)G_k$ and $P(q)$ is the (∞ length) (monic) multivariate prediction error filter for the vector signal G_k and N_k is resulting white prediction error with covariance matrix R_{NN}

$$I - F(q) = \sigma_v^2 R^{-1} R_{NN}^{-1} P(q)\,, I - F_o(q) = \sigma_v^2 R^{-1} R_{NN}^{-1} \tag{5.25}$$

then

$$\begin{aligned} S_{HH} &= \sigma_v^4 R^{-1} R_{NN}^{-1} \int P S_{GG} P^H R_{NN}^{-1} R^{-1} df \\ &\quad + \sigma_v^2 R^{-1} - 2\sigma_v^4 R^{-1} R_{NN}^{-1} R^{-1} \end{aligned} \tag{5.26}$$

$$\begin{aligned} EMSE &= tr(R S_{HH}) \\ &= \sigma_v^2 N - \sigma_v^4 tr\{R_{NN}^{-1} R^{-1}\} \end{aligned} \tag{5.27}$$

$$\begin{aligned} r_{i,NN} &= \exp[\int \ln(s_{h_i h_i}(f) + \frac{\sigma_v^2}{\sigma_y^2})] \\ &= \exp[f_i \ln(\frac{D_i}{f_i} + \frac{\sigma_v^2}{\sigma_y^2}) + (1 - f_i) \ln \frac{\sigma_v^2}{\sigma_y^2}] \\ &= \frac{\sigma_v^2}{\sigma_y^2}\big(1 + \frac{D_i \sigma_y^2}{f_i \sigma_v^2}\big)^{f_i} \end{aligned} \tag{5.28}$$

finally

$$\begin{aligned} EMSE &= \sigma_v^2 N - \sigma_v^4 tr\{R_{NN}^{-1} R^{-1}\} \\ &= \sigma_v^2 N - \sigma_v^4 \sum_{i=1}^{N} \big(1 + \frac{D_i \sigma_y^2}{f_i \sigma_v^2}\big)^{-f_i} \\ &= \sigma_v^2 \sum_{i=1}^{N} \big(1 - \big(1 + \frac{D_i \sigma_y^2}{f_i \sigma_v^2}\big)^{-f_i}\big) \end{aligned} \tag{5.29}$$

EMSE for a non-causal WF

The z-transform of the causal WF window is given by:

$$F(q) = I - S_{\tilde{G}\tilde{G}}(q) S_{GG} = \sigma_v^2 R^{-1} (S_{HH} + \sigma_v^2 R^{-1}) \tag{5.30}$$

$$\begin{aligned} (I - F) S_{HH} (I - F)^H + F S_{\tilde{G}\tilde{G}} F^H &= \sigma_v^4 R^{-1} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} S_{HH} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} R^{-1} \\ &\quad + S_{HH} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} \sigma_v^2 R^{-1} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} S_{HH}^H \end{aligned}$$

then

$$\begin{aligned} tr\{R\big((I - F) S_{HH} (I - F)^H + F S_{\tilde{G}\tilde{G}} F^H\big)\} &= \sigma_v^4 tr\{R^{-1} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} S_{HH} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1}\} \\ &\quad + \sigma_v^2 tr\{S_{HH} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} \big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} S_{HH}^H\} \\ &= \sigma_v^2 tr\{\big(S_{HH} + \sigma_v^2 R^{-1}\big)^{-1} S_{HH}\} \end{aligned}$$

finally

$$EMSE_{ncc} = \sum_{i=1}^{N} \frac{1}{j2\pi} \oint \frac{dz}{z} (\frac{1}{\sigma_v^2} + \frac{1}{\sigma_x^2} D_i S_{hh}(z))^{-1} = \sigma_v^2 2 f_i \sum_{i=1}^{N} \frac{1}{1 + \frac{\sigma_v^2}{\sigma_x^2} D_{ii} \frac{1}{2f_i}} \quad (5.31)$$

EMSE for RLS

The z-transform of the exponential window RLS is given by:

$$F_E(z) = \frac{1-\lambda}{1-\lambda z^{-1}} \quad (5.32)$$

In order to perform the analytical expression of the Excess MSE in the case of a exponential window, we start calculating the quantity $\int_0^{f_i} \left|1 - F_E(e^{2j\pi f})\right|^2 df$.

$$\left|1 - F_E(e^{2j\pi f})\right|^2 = \frac{2\lambda^2 (1 - \cos(2\pi f))}{1 - 2\lambda \cos(2\pi f) + \lambda^2}$$

The previous expression can be expanded as

$$\left|1 - F_E(e^{2j\pi f})\right|^2 = \frac{2\lambda^2 - 2\lambda}{1 - 2\lambda \cos(2\pi f) + \lambda^2} + \frac{2\lambda (1 - \lambda \cos(2\pi f))}{1 - 2\lambda \cos(2\pi f) + \lambda^2} \quad (5.33)$$

On the other hand, we have:

$$\begin{aligned} \int \frac{1-a^2}{1 - 2a\cos(x) + a^2} dx &= 2\text{arctan}\left(\frac{1+a}{1-a} \tan\left(\frac{x}{2}\right)\right) \\ \int \frac{(1 - a\cos x)}{1 - 2a\cos x + a^2} dx &= \frac{x}{2} + \text{arctan}\left(\frac{1+a}{1-a} \tan\left(\frac{x}{2}\right)\right) \end{aligned}$$

Using the integration change of variables $x = 2\pi f$, we have

$$\int_0^{f_i} \left|1 - F_E(e^{2j\pi f})\right|^2 df = \lambda f_i - \frac{\lambda}{\pi} \frac{1-\lambda}{1+\lambda} \text{arctan}\left(\frac{1+\lambda}{1-\lambda} \tan(\pi f_i)\right)$$

Using an exponential windowing for RLS, The Excess MSE is given by:

$$EMSE = N\sigma_n^2 \frac{1-\lambda}{1+\lambda} + 2\sigma_x^2 \sum_{i=1} N D_i \left(\lambda f_i - \frac{\lambda}{\pi} \frac{1-\lambda}{1+\lambda} \text{arctan}\left(\frac{1+\lambda}{1-\lambda} \tan(\pi f_i)\right)\right)$$

Chapter 6

Bayesian Adaptive Filtering : EM-Kalman Algorithm

In this chapter we continue our study of the Bayesian Adaptive Filtering (BAF) concept that we introduced in the previous chapter. The proposed technique is based on modeling the optimal adaptive filter coefficients as a stationary vector process, in particular as a AR(1) model. Optimal adaptive filtering with such a state model becomes Kalman filtering. The complexity of the resulting algorithm is $O(N^3)$ and in order to reduce this complexity we propose a diagonal AR(1) based approach of complexity $O(N^2)$ which is comparable to RLS complexity. For the AR(1) model parameters estimation, we propose an adaptive version of the EM algorithm with complexity limited to $O(N)$. The proposed parameters estimation method leads to linear prediction on reconstructed optimal filter correlations, and hence a meaningful approximation/estimation compromise. To further reduce the initial adaptive EM-Kalman algorithm complexity, we develop a second approach based on component-wise EM-Kalman (This technique is of complexity $O(N)$ which is comparable to LMS complexity). To compare the proposed algorithms performance, we derived the analytical expressions of EMSE in the steady-state in the general case and we proposed a comparison for the application to radio mobile communications where the priori information is the fading, PDP and the Doppler shift. The former proposed algorithm is outperformed by the adaptive EM-Kalman based algorithm, in terms of tracking and convergence. To offer comparable performance with the adaptive

EM-Kalman algorithm with same complexity of component-wise EM, we propose in the following chapter a two-stage technique.

6.1 Parameter Estimation via the EM algorithm

Consider now the prototype adaptive filtering set-up, which is the system identification set-up, in which the desired response signal y_k is modeled as the output of the optimal filter, which can be time-varying, plus independent (white) noise. The adaptive system identification Fig. 6.1 is designed for determining a (typically linear FIR) model of the transfer function for an unknown, time-varying digital or analog system. Let consider

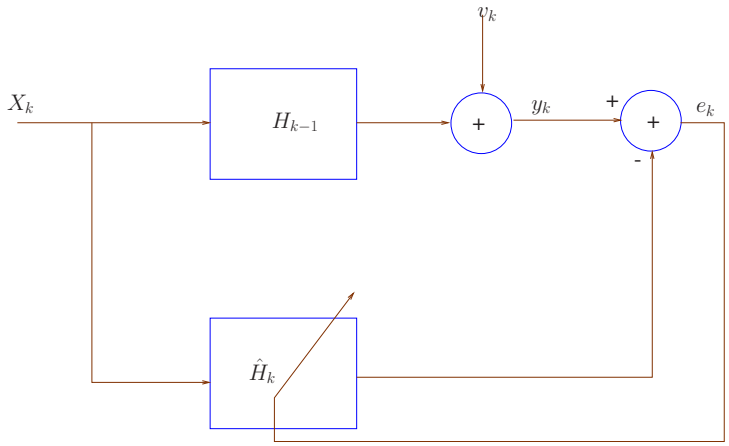

Figure 6.1: System identification block diagram

the system defined bellow

$$\begin{aligned} H_k &= AH_{k-1} + W_k \\ y_k &= X_k^H H_{k-1} + v_k \end{aligned} \tag{6.1}$$

In this section we develop the EM algorithm for estimating the parameters of (6.1). Perhaps the most important step in applying the EM algorithm presented in the previous chapter, to a particular problem is choosing the missing data. The missing data should be chosen so that the maximization of $U(\theta, \theta^k)$ for any value of $\theta = (A, Q)$ is easy to perform, and that the expectation step is possible .

Fortunately, in this case, the choice of missing data is not too difficult.
Let us imagine for a moment that, in addition to the system inputs and outputs, H_k and Y_k respectively, the state H_k was available thus, ML estimation of A reduces to its application to (6.1). The covariance elements, Q, of W_k could then be calculated from the residuals. Moreover, the conditional expectation of state sequence is calculated using a (slightly augmented) Kalman Smoother. All of this suggests that the state sequence is a desirable conditionat for the missing data. We therefore designate Y as the incomplete data so that the complete data set is $Z = (H_k, Y_k)$.
In order to develop a procedure for estimating the parameters in the state-space model defined by (6.1), we note first that the joint log-likelihood of the complet data Z can be written in the form
First, by repeated application of Bays Rule

$$f_Z(z, \theta) = f(H|Y = y).f_Y(y; \theta) \tag{6.2}$$

where $f_Z(z, \theta)$ is the probability density associated with Z and $f_{Z|Y=y}(z, \theta).f_Y(y; \theta)$ is the conditional probability density of Z given $Y = y$. Taking the logarithm on both sides of (6.2),

$$\log f_Y(y, \theta) = \log f_Z(z, \theta) - \log f(H|Y = y) \tag{6.3}$$

Note that the logarithm function is monotonic in its semi-positive argument and any probability density function (p.d.f.) is semi-positive, it follows that the maximising argument of any p.d.f. will be the same as for the logarithm of that function.
Of course equation (6.3) requires knowledge of the complete data set and therefore cannot be calculated. Suppose that, instead of calculating equations (6.3), we calculate an approximation of (6.3) derived as an expectation over the space of H_N, and conditioned upon the actual observations, as well as some estimate of the vector θ say $\hat{\theta}$ then we obtain

$$E_{\hat{\theta}}\{log f_Y(y, \theta) \mid Y\} = E_{\hat{\theta}}\{\log f_Z(z, \theta) \mid Y\} - E_{\hat{\theta}}\{\log f(H|Y = y) \mid Y\} \tag{6.4}$$

or alternatively,

$$L(\theta) = U(\theta, \hat{\theta}_k) - V(\theta, \hat{\theta}_k)$$

where the following definitions have been used.

$$\begin{aligned} L(\theta) &= \log f_Y(y,\theta) \\ U(\theta,\hat{\theta}_k) &= E_{\hat{\theta}}\{\log f_Z(z,\theta) \mid Y\} \\ V(\theta,\hat{\theta}_k) &= E_{\hat{\theta}}\{\log f(H|Y=y) \mid Y\} \end{aligned}$$

We can interpret the function $U(\theta,\hat{\theta}_k)$ as the projection of the likelihood function that we want to solve onto the space spanned by Z and in directions informed by $\hat{\theta}$. In other words, it is our estimate of the log-likelihood function associated with the complet data.

With this definition we can writ

$$\begin{aligned} L &= \log f_\theta(H_k, Y_M, \theta|Y_M) \\ &= -M\log detQ - M\log detR + \sum_{k=1}^{M} tr(H_k - AH_{k-1})Q^{-1}(H_k - AH_{k-1})^H \\ &\quad + \sum_{k=1}^{M} tr(y_k - X_k^H H_{k-1})R^{-1}(y_k - X_k^H H_{k-1})^H \end{aligned} \tag{6.5}$$

The log-likelihood given above depends on the unobserved data H_k. We consider applying the EM algorithm conditionally with respect to the observed ensemble Y. That is, the estimated parameters at the $(k+1)-th$ iteration are the values A and Q that maximize

$$U(\theta,\hat{\theta}_k) = E_{\hat{\theta}_k}\{\log f_\theta(H_k, Y_M, \theta|Y_M)\} \tag{6.6}$$

where $E_{\hat{\theta}_k}$ denotes the conditional expectation relative to a density containing the $k-th$ iteration values.
In order to calculate the conditional expectation defined in (**??**), it is convenient to define the conditional mean

$$\begin{aligned} \hat{H}_k &= E_{\hat{\theta}_k}\{H_k^0|Y_M\} \\ P_k &= E[\tilde{H}_k\tilde{H}_k^H|Y_M\} \\ P_{k-1} &= E[\tilde{H}_{k-1}\tilde{H}_{k-1}^H|Y_M\} \end{aligned}$$

we suppose the following definitions

$$\begin{aligned}
\mathbf{\Pi_{k|k}} &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_{k-1}(H_{k-1})^H|Y_M\} + P_{k-1}) \\
\mathbf{\Pi_{k-1|k}} &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_k H_k^H|Y_M\} + P_k) \\
\mathbf{\Pi_{k,k-1|k}} &= \sum_{k=1}^{M}(E_{\hat{\theta}_k}\{H_k H_{k-1}^H|Y_M\} + P_{k,k-1})
\end{aligned} \tag{6.7}$$

The Kalman filter terms $\hat{H}_k$, P_k and $P_{k,k-1}$ are computed under the parameter values A_k and Q_k using the recursions in (6.7). Maximizing (6.6) w.r.t. A and Q, we obtain **Maximum with respect to** a: Differentiating the expected log-likelihood with respect to A yields:

$$\frac{\partial E[l(\theta)]}{\partial A} = Q^{-1}\frac{-1}{2}(-2\Pi_{k,k-1} + 2A\Pi_{k-1}) \tag{6.8}$$

Equating this result to zero yields the value of A that maximizes the approximate log-likelihood:

$$\mathbf{A_{k+1}} = \mathbf{\Pi_{k,k-1|k}}(\mathbf{\Pi_{k-1|k}})^{-1} \tag{6.9}$$

Maximum with respect to R: Differentiating the expected log-likelihood with respect to R^{-1} gives:

$$\frac{\partial E[l(\theta)]}{\partial R^{-1}} = -\sum_{k=1}^{M}\frac{1}{2}(y_k - \hat{y}_k)(y_k - \hat{y}_k)^* + \frac{M}{2}R \tag{6.10}$$

Hence, by equating the above result to zero, the maximum of the approximate log-likelihood with respect to R is given by:

$$R = \frac{1}{M}\sum_{k=1}^{M}(y_k - \hat{y}_k)(y_k - \hat{y}_k)^* \tag{6.11}$$

Maximum with respect to Q: Maximum with respect to q Following the same steps, the derivative of the expected log-likelihood with respect to Q^{-1} is given by:

$$\frac{\partial E[l(\theta)]}{\partial R^{-1}} = -\frac{1}{2}(\Pi_k - 2A\Pi_{k,k-1}^* + A\Pi_{k-1}A^*) + \frac{M}{2}Q$$

Hence, equating to zero and using the result that
$\mathbf{A_{k+1}} = \mathbf{\Pi_{k,k-1|k}}(\mathbf{\Pi_{k-1|k}})^{-\mathbf{1}}$, the maximum of the approximate log-likelihood with respect to q is given by:

$$\mathbf{Q_{k+1}} = \frac{1}{M}(\mathbf{\Pi_{k|k}} - \mathbf{\Pi_{k,k-1|k}}(\mathbf{\Pi_{k-1|k}})^{-\mathbf{1}}\mathbf{\Pi^{H}_{k,k-1|k}}) \tag{6.12}$$

6.2 Adaptive EM-Kalman Algorithm

In our study, the tasks of smoothing in a missing data context, introduced in the previous chapter, are interpreted as basically the problem of estimating the BAF H_k in the state-space model (6.1). The conditional means provide a minimum MSE solution based on the observed data. The parameters Q and A are estimated using the EM algorithm. The estimation of the optimal filter variation is carried out by KF'ing and one step smoothing and we introduce an EM approach to iteratively update the parameter model.
The resulting algorithm is shown in Table 6.2 of Adaptive EM-Kalman filter. The complexity of Kalman filter is limited to $O(N^2)$ order and the Adaptive Kalman filter has the same order of complexity. To reduce the complexity of our algorithm we propose a Component-Wise Adaptive Kalman filter, which is based on the estimation of each parameter one by one.

6.3 MAP-ML Estimation

The value of H_k that maximizes the posterior density (that is, the mode of the posterior density) is called the maximum a posterior probability estimate of H_k.
If the posterior density of H_k given A, Q and Y is unimodal and symmetric, then it is easy to see that the MAP estimate and the mean squared estimate coincide, since the posterior density attains its maximum value at its expected value.

Adaptive EM-Kalman Algorithm	
Computation	**Cost** $(\times)$
Initialization	
$\widehat{\mathbf{H}}_{\mathbf{0|0}} = \hat{\mathbf{0}}$, $\mathbf{P_{0|0}} = \mathbf{100I}$, $\mathbf{A_0} = \alpha\mathbf{I}$, $\mathbf{Q_0} = (\mathbf{1}-\alpha)\mathbf{I}$ $\mathbf{\Pi_{0|0}} = \mathbf{0}$, $\mathbf{\Pi_{1,0|0}} = \mathbf{0}$, $\mathbf{\Pi_{1|0}} = \mathbf{0}$ $\gamma^{(0)} = 0$	$2N$
Kalman filtering and one step smoothing	
$\widehat{\mathbf{H}}_{\mathbf{k|k-1}} = \hat{\mathbf{A}}_{\mathbf{k}}\widehat{\mathbf{H}}_{\mathbf{k-1|k-1}}$	N^2
$\hat{y}_{k\|k-1} = \mathbf{x}_{\mathbf{k}}^{\mathbf{H}}\widehat{\mathbf{H}}_{\mathbf{k|k-1}}$	N^2
$\mathbf{K_k} = \mathbf{P_{k|k-1}x_k}$	N^2
$\mathbf{M_k} = (\mathbf{x_k^H K_k} + \sigma_{\mathbf{v}}^{\mathbf{2}})^{-\mathbf{1}}$	N
$\mathbf{K_k^f} = \mathbf{K_k M_k}$	1
$\mathbf{C_{k-1}} = \mathbf{P_{k-1|k-1}A_k^H P_{k|k-1}^{-1}}$	$2N^2$
$\widehat{\mathbf{H}}_{\mathbf{k-1|k}} = \widehat{\mathbf{H}}_{\mathbf{k|k-1}} + \mathbf{K_k^f}(\mathbf{y_k} - \hat{\mathbf{y}}_{\mathbf{k|k-1}})$	1
$\mathbf{P_{k|k-1}} = \mathbf{A_k P_{k-1|k-1} A_k^H} + \mathbf{Q_k}$	$N(\frac{N-1}{2})$
$\widehat{\mathbf{H}}_{\mathbf{k|k}} = \widehat{\mathbf{H}}_{\mathbf{k|k-1}} + \mathbf{A_k^{-1}}(\mathbf{K_k} - \mathbf{Q_k x_k})\mathbf{M_k}(\mathbf{y_k} - \hat{\mathbf{y}}_{\mathbf{k|k-1}})$	$N+1$
$\mathbf{P_{k|k}} = \mathbf{P_{k|k-1}} - \mathbf{K_k^f K_k^H}$	1
$\mathbf{P_{k-1|k}} = \mathbf{P_{k-1|k-1}} + \mathbf{C_{k-1}}(\mathbf{P_{k|k}} - \mathbf{P_{k|k-1}})$	N^2
Model Parameters Adaptation	
$\mathbf{\Pi_{k|k}} = \lambda\mathbf{\Pi_{k|k-1}} + \mathbf{diag}(\widehat{\mathbf{H}}_{\mathbf{k|k}}\widehat{\mathbf{H}}_{\mathbf{k|k}}^{\mathbf{H}} + \mathbf{P_{k|k}})$	N
$\mathbf{\Pi_{k-1|k}} = \lambda\mathbf{\Pi_{k-1|k-1}} + \mathbf{diag}(\widehat{\mathbf{H}}_{\mathbf{k-1|k}}\widehat{\mathbf{H}}_{\mathbf{k-1|k}}^{\mathbf{H}} + \mathbf{P_{k-1|k}})$	N
$\mathbf{D_k} = \mathbf{P_{k|k}C_{k-1}^H} = \mathbf{A_k P_{k-1|k-1}} - \mathbf{K_k^f}(\mathbf{A_k^{-1}}(\mathbf{K_k} - \mathbf{Q_k x_k}))^{\mathbf{H}}$	$2N$
$\mathbf{\Pi_{k,k-1|k}} = \lambda\mathbf{\Pi_{k,k-1|k-1}} + \mathbf{diag}(\widehat{\mathbf{H}}_{\mathbf{k|k}}\widehat{\mathbf{H}}_{\mathbf{k-1|k}}^{\mathbf{H}} + \mathbf{D_k})$	N
$\mathbf{Q_{k+1}} = \frac{\mathbf{1}}{\gamma_{\mathbf{k}}}(\mathbf{\Pi_{k|k}} - \mathbf{\Pi_{k,k-1|k}}(\mathbf{\Pi_{k-1|k}})^{-\mathbf{1}}(\mathbf{\Pi_{k,k-1|k}})^{\mathbf{H}})$	$2N$
$\gamma_k = \gamma_{k-1} + 1$	
$\mathbf{A_{k+1}} = \mathbf{\Pi_{k,k-1|k}}(\mathbf{\Pi_{k-1|k}})^{-\mathbf{1}}$	N
cost/update $7.5N^2 + 11.5N + 3$	

Figure 6.2: Adaptive EM-Kalman Algorithm

Let the sequence filter H_k be considered as a random variable distributed according to the posterior density $f_{H_k}(h_k)$. The posterior distribution for H, is given by

$$f_{H_k,Y|A,Q}(h_k, y \mid A, Q) \tag{6.13}$$

then $\widehat{H_{kMAP}}$ is obtained by maximizing the logarithm of the posterior density with respect to H_k. Initially, A_0 and Q_0 are set to a certain initial value. After the first iteration, A_{k+1} and Q_{k+1} are obtained by ML, given $\widehat{H}_{k,MAP}$.

6.4 Component-wise Adaptive Kalman Algorithm

Our goal is to design an optimal algorithm with reduced complexity in a realistic environment, considering the filter coefficients to estimate as random variables. In a previous section, a Bayesian Adaptive Filtering (BAF) approach has been proposed, showing a complexity of order $O(N^2)$. To reduce the complexity of the algorithm presented in Table of Adaptive EM-Kalman filter, we propose a Component-Wise Adaptive Kalman algorithm to update the filter coefficients, which decreases computational complexity in 1 order of magnitude while preserving convergence. Experimental results will be shown for the proposed algorithm, comparing to KF filtering and Adaptive Kalman algorithms.The filter parameters are iteratively computed through M iterations. The system (6.1 becomes for $n = 1 \dots N$, where N is the length of the filter

$$h_{k,n} = a_n h_{k-1,n} + w_{k,n} \tag{6.14}$$

$$y_k = h_{k-1,n} x_{k,n} + \sum_{j \neq n}^{N} h_{k-1,n} x_{k,n} + v_k \tag{6.15}$$

and

$$h_k = \hat{h}_k + \tilde{h}_k$$

we can write

$$y_k - \sum_{j \neq n}^{N} \hat{h}_{k-1,n} x_{k,n} = h_{k-1,n} x_{k,n} + \sum_{j \neq n}^{N} \tilde{h}_{k-1,n} x_{k,n} + v_k$$

In each iteration y_k and v_k are updated as follows

$$y'_k = y_k - \sum_{j \neq n}^{N} \hat{h}_{k-1,n} x_{k,n}$$

and

$$v'_k = \sum_{j \neq n}^{N} \tilde{h}_{k-1,n} x_{k,n} + v_k$$

Component Wise EM-Kalman algorithm
- Begin with an arbitrary set
- For i=1:N -Compute: $h_{i,k} = a_i h_{i,k-1} + w_{i,k}$ $y_k = x^*_{i,k} h_{i,k-1} + \sum_{j=1}^{i-1} x^*_{i,k} h_{i,k-1} + v_k$ $= x^*_{i,k} h_{i,k-1} + \sum_{j=1}^{i-1} x^*_{i,k} (\hat{h}_{i,k-1} - \tilde{h}_{i,k-1}) + v_k$ $\Longrightarrow$ $y_k - \sum_{j=1}^{i-1} x^*_{i,k} (\hat{h}_{i,k-1}) = x^*_{i,k} h_{i,k-1} - \sum_{j=1}^{i-1} x^*_{i,k} \tilde{h}_{i,k-1} + v_k$ $y'_k = x^*_{i,k} h_{i,k-1} + v'_k$- EM-Kalman -Complexilty O(N) order

The missing data should be chosen so that the task of maximizing $U(\theta, \theta_n^k)$, for $n = 1 \ldots N$, $\theta_n^l = (a_n, q_n)$ is easy and so that it is possible to perform the expectation step. Fortunately, in this case, the choice of missing data is not too difficult. Let us imagine for a moment that, in addition to the system inputs and outputs, $x_{k,n}$ and Y_k respectively, the state h_n was available then ML estimation of a_n reduces to applying to (6.15). The covariance elements, q_n, of w_k could then be calculated from the residuals. Moreover, the conditional expectation of state sequence may be calculated using a (slightly augmented) Kalman Smoother. All of this suggests that the state sequence is a desirable conditionat for the missing data. We therefore designate Y as the incomplete data so that the complete data set is $Z = (h_n, Y)$.

For the $n-th$ iteration the log-likelihood function can be written as

$$\begin{aligned} L &= -2\log f_\theta(h_n, Y_M, \theta|Y) \\ &= N\log detq_n \\ &\quad + \sum_{k=1}^{M} q_n^{-2}(h_{k,n} - a_n h_{k-1,n})(h_{k,n} - a_n h_{k-1,n})^H \\ &\quad + \sum_{k=1}^{M} \sigma_v^{-2}(y_k - x_{k,n}^H h_{k-1,n})(y_k - x_k^H h_{k-1,n})^H \end{aligned} \tag{6.16}$$

where L is to be maximized with respect to parameters a_n and q_n . Since the log-likelihood given above depends on the unobserved data $h_{k,n}$, we consider applying the EM algorithm conditionally with respect to the observed Y. That is, the estimated parameters at the $(k+1)-th$ iterate as the values a_n and q_n which maximize

$$U(\theta, \hat{\theta}_k) = E_{\hat{\theta}_k}\{\log f_\theta(h_n, Y_M, \theta|Y)\} \tag{6.17}$$

where $E_{\hat{\theta}_k}$ denotes the conditional expectaion relative to a density containing the $k\,th$ iterate values.

In order to calculate the conditional expectation defined in (6.5), it is convenient to define the conditional mean

$$\hat{h}_{k,n} = E_{\hat{\theta}_k}\{h_{k,n}|Y\}$$

and

$$P_{k,n} = E\{\tilde{h}_{k,n}\tilde{h}_{k,n}^H\}$$

we suppose the following definitions

$$\begin{aligned} \pi_{k,n|k} &= \sum_{k=1}^{M} E_{\hat{\theta}_k}\{h_{k-1,n}h_{k-1,n}^H|Y\} + P_{k-1,n} \\ \pi_{k-1,n|k} &= \sum_{k=1}^{M} E_{\hat{\theta}_k}\{h_{k,n}h_{k,n}^{oH}|Y\} + P_{k,n} \\ \pi_{k,k-1,n|k} &= \sum_{k=1}^{M} E_{\hat{\theta}_k}\{h_{k,n}h_{k-1,n}^H|Y\} + P_{k,k-1} \end{aligned} \tag{6.18}$$

The Kalman filter terms $\hat{h}_{k,n}$, $P_{k,n}$ and $P_{k,k-1}$ are computed under the parameter values $a_{n,k}$ and $q_{n,k}$ using the recursions in (6.16).Furthermore, it is easy to see that the choices

$$\begin{aligned} q_{k+1,n+1} &= \frac{1}{\gamma_k}(\pi_{\mathbf{k|k}} - \pi_{\mathbf{k,k-1|k}}(\pi_{\mathbf{k-1|k}})^{-\mathbf{1}}(\pi_{\mathbf{k,k-1|k}})^{\mathbf{H}}) \\ a_{k+1,n+1} &= \pi_{\mathbf{k,k-1|k}}(\pi_{\mathbf{k-1|k}})^{-\mathbf{1}} \end{aligned}$$

maximize the last two lines in the Expectation-likelihood function (6.5). In our study, the tasks of smoothing in a missing data context are interpreted as basically the problem of estimating the BAF $h_{k,n}$ in the state-space model (6.15). The conditional means provide a minimum MSE solution based on the observed data. The parameters q_n and a_n are estimated by ML using the component-wise EM algorithm. We simplify the estimation problem by considering a_n and q_n diagonal matrices. The filter parameters are iteratively computed through M iterations. The estimation of the optimal filter variation is carried out by KF'ing and one step smoothing and we introduce an EM approach for iteratively update the parameter model.

The algorithm is resulting in Table of component-wise Adaptive EM-Kalman.

Adaptive Component-Wise EM-Kalman Algorithm	
Computation	**Cost** $(\times)$
Initialization	
$\hat{h}_{0\|0} = 0\ ,\ \ P_{0\|0} = 100\ ,$	
$a_0 = \alpha\ ,\ q_0 = (1-\alpha)$	
$\pi_{0\|0} = 0\ ,\ \ \pi_{1,0\|0} = 0,\ \ \pi_{1\|0} = 0$	
$\gamma_0 = 0$	
Kalman filtering and one step smoothing	
$for\ n = 1 \dots N$	
$\hat{h}_{k,n\|k-1} = a_{n,k}\hat{h}_{k-1,n\|k-1}$	1
$\hat{y}_{k,n\|k-1} = x^H_{n,k}\hat{h}_{k,n\|k-1}$	1
$K_{n,k} = P_{k,n\|k-1}x_k$	1
$M_{k,n} = (x^H_{k,n}K_{n,k} + \sigma^2_v)^{-1}$	1
$K^f_{n,k} = K_{n,k}M_{n,k}$	1
$C_{n,k-1} = P_{n,k-1\|k-1}a^H_{n,k}P^{-1}_{n,k\|k-1}$	2
$\hat{h}_{n,k-1\|k} = \hat{h}_{n,k\|k-1} + K^f_{n,k}(y_k - \hat{y}_{k,n\|k-1})$	1
$P_{k,n\|k-1} = a_{n,k}P_{k-1,n\|k-1}a^H_{n,k} + q_{n,k}$	1
$\hat{h}_{k,n\|k} = \hat{h}_{k,n\|k-1} + a_{n,k}(K_k - q_{n,k}x_k)M_k(y_k - \hat{y}_{k,n\|k-1})$	2
$P_{k,n\|k} = P_{k,n\|k-1} - K^f_{n,k}K^H_{n,k}$	1
$P_{k-1,n\|k} = P_{k-1,n\|k-1} + C_{n,k-1}(P_{n,k\|k} - P_{n,k\|k-1})$	1
Model Parameters Adaptation	
$\pi_{n,k\|k} = \lambda\pi_{k,n\|k-1} + diag(\hat{h}_{n,k\|k}\hat{h}^H_{n,k\|k} + P_{n,k\|k})$	1
$\Pi_{n,k-1\|k} = \lambda\pi_{n,k-1\|k-1} + diag(\hat{h}_{n,k-1\|k}\hat{h}^H_{n,k-1\|k} + P_{n,k-1\|k})$	1
$D_{n,k} = P_{n,k\|k}C^H_{n,k-1} = a_{n,k}P_{n,k-1\|k-1} - K^f_{n,k}(a^{-1}_{n,k}(K_{n,k} - q_{n,k}x_{n,k}))^H$	2
$\pi_{n,k,k-1\|k} = \lambda\pi_{n,k,k-1\|k-1} + diag(\hat{h}_{n,k\|k}\hat{H}^H_{n,k-1\|k} + D_{n,k})$	1
$q_{n,k+1} = \frac{1}{\gamma_k}(\pi_{n,k,\|k} - \pi_{n,k,k-1\|k}(\pi_{n,k-1\|k})^{-1}(\pi_{n,k,k-1\|k})^H)$	2
$\gamma_k = \gamma_{k-1} + 1$	
$a_{n,k+1} = \pi_{n,k,k-1\|k}(\pi_{n,k-1\|k})^{-1}$	1
cost/update $21N$	

6.5 Simplified Component-Wise adaptive Kalman algorithm

We consider A to be an identity matrix. Hence the complexity of the adaptive part is comparable to the one exhibited by tap Variable Step-Size (TVSS) LMS [21], [17, 18, 14], like $4N$. In practice A tends to the identity matrix when MSE converges to $MMSE$. The process is low-pass which is equivalent to a random walk.

6.6 Performance Analysis

In this section we will compare the performance of the BAF and SAF in terms of the resulting EMSE.

6.6.1 Steady-State Excess Mean-Square Error (EMSE)

The state estimate update is given by Kalman as :

$$\begin{aligned}
\hat{H}_{k|k} &= A\hat{H}_{k-1|k-1} + K_k(y_k - X_k A\hat{H}_{k-1|k-1}) \\
&= A\hat{H}_{k-1|k-1} + K_k X^H(H_{k-1} - A\hat{H}_{k-1|k-1}) \\
&\quad + K_k e_{opt} \\
&= A\hat{H}_{k-1|k-1} \\
&\quad + \frac{P_{k|k}X_k X_k^H}{\sigma_v^2}(H_{k-1} - A\hat{H}_{k-1|k-1}) \\
&\quad + K_k e_{opt}
\end{aligned} \tag{6.19}$$

Where $K_k = \frac{P_{k|k}X_k}{\sigma_v^2}$

and e_{opt} represents the minimum(in a mean square sense) error at time k. In studying tracking behavior, we may exclude the influence of the estimation noise, since the deviation of $E[H_{k|k}]$ from H_k determines the response of the BAF algorithm to the non-

stationarity of the environment. Taking expected values on both sides of (7.20), we get

$$\begin{aligned} E[\hat{H}_{k|k}] &= AE[\hat{H}_{k-1|k-1}] \\ &\quad + \frac{P_{k|k}E[X_k X_k^H]}{\sigma_v^2})(H_{k-1} - AE[\hat{H}_{k-1|k-1}]) \\ &= AE[\hat{H}_{k-1|k-1}] \\ &\quad + \frac{P_{k|k}R}{\sigma_v^2}(H_{k-1} - AE[\hat{H}_{k-1|k-1}]) \end{aligned} \tag{6.20}$$

the lag-error is given by

$$\tilde{H}_k = E[\hat{H}_{k|k}] - H_k$$

$$\begin{aligned} \tilde{H}_k &= AE[\hat{H}_{k-1|k-1}] - \frac{P_{k|k}RA}{\sigma_v^2}(\tilde{H}_{k-1}) \\ &\quad + \frac{P_{k|k}R}{\sigma_v^2}(H_{k-1} - AH_{k-1}) - H_k + AH_{k-1} \\ &= AE[\hat{H}_{k-1|k-1}] - \frac{P_{k|k}RA}{\sigma_v^2}(\tilde{H}_{k-1}) \\ &\quad + (\frac{P_{k|k}R}{\sigma_v^2}(I - A) + A)H_{k-1} - H_k \\ &\approx (I - \frac{P_{k|k}R}{\sigma_v^2})A\tilde{H}_{k-1} + AH_{k-1} - H_k \end{aligned} \tag{6.21}$$

The input is considered to be white ($R = \sigma_x^2 I$), note that, each element of the lag-error vector is determined by the following relation:

$$\tilde{h}_i(k) = (1 - \frac{p_{i,k|k}\sigma_x^2}{\sigma_v^2})A_i\tilde{h}_{i,k-1} + A_i h_{i,k-1} - h_{i,k}. \tag{6.22}$$

where $\tilde{h}_i(k)$ is the i^{th} element of $\tilde{H}_k$. By properly interpreting the equation above, we can say that the lag is generated by applying the transformed instantaneous optimal coefficient to a first-order discrete-time filter denoted ***lag filter***

$$\tilde{H}_i(z) = \frac{A_i z^{-1} - 1}{1 - (1 - \frac{\sigma_x^2 p_{i,k|k}}{\sigma_v^2})A_i z^{-1}} H_i(z) \tag{6.23}$$

Using the inverse z-transform, the variance of the elements of the vector $\tilde{H}(k)$ can then be calculated by

$$E[\tilde{h}_i(k)\tilde{h}_i^H(k)] = \frac{1}{2\pi j}\oint \tilde{H}_i(z)\tilde{H}_i(z^{-1})Q_i z^{-1} dz$$

The BAF excess mean square error due to lag is then given by Eq. (7.25)

$$EMSE = \sum_{i=1}^{N} EMSE_i \tag{6.24}$$

where

$$EMSE_i = \sigma_x^2 p_i = \sigma_x^2 \frac{-(\frac{Q_i}{SNR_i} + Q_i) + Q_i\sqrt{(1+\frac{1}{SNR_i})^2 + 4\frac{\sigma_x^2}{\sigma_v^2} \mid A_i \mid^2}}{2 \mid A_i \mid^2} \tag{6.25}$$

where $SNR_i = \frac{\sigma_x^2 Q_i}{\sigma_v^2(1-|A_i|^2)}$, the time constant is given by $\tau = \frac{1}{1-|A_i|}$ and $Q_i = Q_1\beta^i$ (where β_i is the PDP).

6.6.2 Simplified Expression for Bayesian Adaptive Filtering

In simplified scenarios (e.g. high SNR, slow variation) we can write

$$\begin{aligned} p_i &= \frac{1- \mid A_i \mid^2}{1+SNR_i} \\ &= \frac{4\pi\beta_i}{1+SNR_i} \end{aligned}$$

where $1- \mid A_i \mid^2 = 4\pi\beta_i$

The misadjustement of BAF is given by

$$\begin{aligned} M &= \frac{EMSE}{MMSE} \\ &= 2\pi\frac{\sigma_x^2}{\sigma_v^2}\sum_{i=1}^{N}\frac{\beta_i}{1+SNR_i} \end{aligned}$$

If all β_i are the same, then $\sum_{i=1}^{N}\frac{1}{1+SNR_i}$ is the inverse of the harmonic mean of $1+SNR_i$ and $\frac{1}{(\sum_{i=1}^{N}\frac{1}{1+SNR_i})} \leq \sum_{i=1}^{N}(1+SNR_i)$ (the arithmetic mean) And β_i can be lows if A_i is complex $A_i = \mid A_i \mid e^{j2\pi f_i}$, where f_i is Doppler shift and $\mid A_i \mid = 1 - 2\pi\beta_i$ (if $\mid A_i \mid$ is sufficiently close to one) Then in this case the EMSE of Kalman, depend only on the PDP β_i, while the EMSE in the SAF case depend on the PDP and f_i.

and the SAF excess mean square error due to lag is then given by Eq. (6.39).

6.7 RLS and LMS EMSE's style BAF

6.7.1 Tracking Behavior

The problem we examine here involves adaptive filtering of an unknown time-varying system using an LMS and RLS transversal filters. Also, it is of interest to learn how the tracking error in $\hat{H}_k$ affects the output MSE ([21]). Here, the effect of the measurement noise are not considered, since only the nonstationary effects are considered. Also, both effects on the MSE can be added since, in general, they are independent.
For an RLS transversal filter, the adaptation equation yields

$$\begin{aligned} \hat{H}_k &= \hat{H}_{k-1} + K_K e_k, \\ &= \hat{H}_{k-1} + \hat{R}_k^{-1} X_k X_k^H \big(H_{k-1} - \hat{H}_{k-1}\big) + K_k v_k \end{aligned} \tag{6.26}$$

where we have defined

$$y_k = X_k^H H_{k-1} + v_k \tag{6.27}$$

where v_k represents the minimum (in a mean square sense) error. In studying behavior, we may exclude the influence of the estimation noise, since the deviation of $E\hat{H}_k$ from H_k determines the response of the adaptive algorithm to the non-stationarity of the environment. Tracking expected values on both sides of (6.26) and assuming that $\hat{R}_k$ is independent of X_k and v_k, and assuming the fact that X_k and v_k are orthogonal, we get

$$E[\hat{H}]_k = E[\hat{H}_{k-1}] + E[\hat{R}_k^{-1} X_k X_k^H]\big(H_{k-1} - E[\hat{H}_{k-1}]\big). \tag{6.28}$$

In ([21]) show that if the variations of the environment are slow with respect to the memory of the adaptive algorithm then, independently of the adaptive system structure, we approximately have

$$E[\hat{R}_k^{-1} X_k X_k^H] = (1-\lambda) I \tag{6.29}$$

Substituting (6.28) into (6.29), we finally get

$$E[\hat{H}]_k = E[\hat{H}_{k-1}] + (1-\lambda)\big(H_{k-1} - E[\hat{H}_{k-1}]\big). \tag{6.30}$$

Now defining the lag-error vector in the coefficients as

$$\tilde{H}_k = E[\hat{H}]_k - H_k \tag{6.31}$$

From equation (6.30) it can be concluded

$$\tilde{H}_k = \lambda \tilde{H}_{k-1} - H_k + H_{k-1}. \tag{6.32}$$

From (6.32) is equivalent to say that the lag is generated by applying the optimal instantaneous value H_k through a first order discrete-time filter as follows:

$$\tilde{H}_i(z) = \frac{z^{-1} - 1}{1 - \lambda z^{-1}} H_i(z) \tag{6.33}$$

The discrete-time filter transient response converges with a time constant given by

$$\tau = \frac{1}{1 - \lambda} \tag{6.34}$$

The time constant is of course the same for each individual coefficient. Note that the tracking ability of the coefficients in the RLS algorithm is independent of the input-signal correlation-matrix eigenvalues.
In the BAF the optimal coefficients values are an AR(1)

$$H_k = AH_{k-1} + W_k \tag{6.35}$$

Where, $E[W_k W_k^H] = Q$. A and Q are diagonal.
The excess mean square error du to lag is then given by

$$\begin{aligned} EMSE &= E[\tilde{H}_k^H R \tilde{H}_k] \\ &= E[tr(R\tilde{H}_k^H \tilde{H}_k)] \\ &= tr(RE[\tilde{H}_k^H \tilde{H}_k]) \end{aligned} \tag{6.36}$$

For the unknown coefficient vector with the model above, the lag-error vector can be generated by applying the elements of the noise process W_k to discrete-time filter with transfer function

$$F(z) = \frac{(z^{-1} - 1)}{(1 - \lambda z^{-1})(1 - A_i z^{-1})} \tag{6.37}$$

Where A_i represent the i^{th} diagonal element of A.

This transfer function consists of a cascade of the lag filter with the all pole filter representing the AR(1). The solution for the variance of the lag terms $\tilde{H}_{i,k}$ can be computed through the inverse z-transform as follows:

$$E[\tilde{H}_{i,k}^H \tilde{H}_{i,k}] = \frac{1}{2\pi j} \oint F(z)F(z^{-1})Q_i z^{-1} dz \tag{6.38}$$

The integral above can be solved using the residue theorem. Assuming the input signal to be white $R = \sigma_x^2$, then the EMSE du to lag is given by

$$EMSE^{RLS} = \sigma_x^2 \sum_{i=1}^{N} \frac{Q_i}{A_i(1+\lambda^2) - \lambda(1+A_i^2)} \left(\frac{1-\lambda}{1+\lambda} - \frac{1-A_i}{1+A_i}\right) \tag{6.39}$$

$$\tag{6.40}$$

With the same analysis we obtain the EMSE for LMS as follows

$$EMSE^{LMS} = \frac{1}{4\mu} \sum_{i=1}^{N} \frac{A_i Q_i}{1 - \sigma_x^2 A_i^2}) \tag{6.41}$$

This result is obtained assuming relatively slow variation:

$$P \approx E[P] \; and \; E[P^{-1}] \approx (E[P])^{-1}$$

The Fig. (6.3) shows, for a slow variation the approximative EMSE is the same one that real EMSE given by recursive Riccati equations.

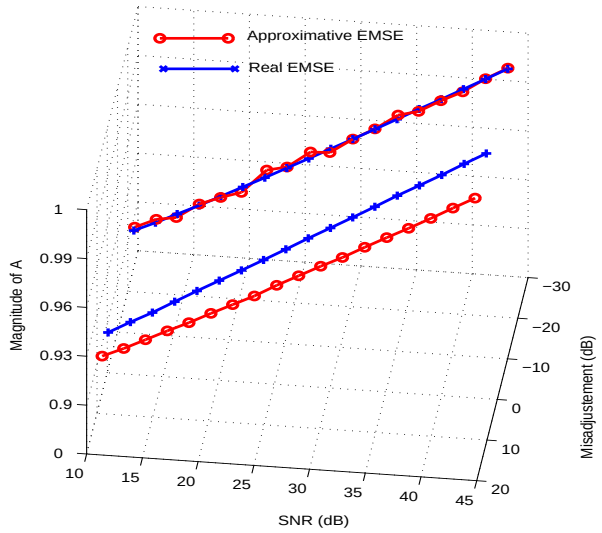

Figure 6.3: Comparative tracking performance results between misadjustement given by exact and approximate EMSEs for slow ($A_i = 0.99\ \tau = 5N$) and Medium ($A_i = 0.90\ \tau = 0.5N$) variations at SNR = 15 dB and $\beta = 0.9$

6.8 Application: Mobile Radio Channel

In this section we use the same application as in the last chapter, with a different structure of the channel model.

The model in this chapter will take into consideration all parameters characterizing the channel without any exception. The matrix A represent the attenuation plus the angle of arrival $A =\mid A \mid e^{-j\phi}$ and Q is the matrix representing the PDP.

6.9 Numerical Results

Here, we consider a non-stationary environment and we compare the behavior of BAF (given by an Adaptive EM-Kalman algorithm) and standard adaptive filters. In all simulations presented here, the desired signal y_k is corrupted by zero mean, (iid) Gaussian noise of variance σ_v^2.

The proposed algorithms are implemented with the model parameters $\mid A_i \mid = 1 - \alpha/N$,

with $\alpha = 0.4$,ϕ is considered random, $Q_i = Q1\beta^i$ is chosen such as $\frac{Q_N}{Q_1} \ll 1$ and the length of the filter is $N = 20$.

Fig. 6.4 and Fig. 6.5 plot the total Excess Mean Square Error ($EMSE$) as a function of the PDP for a BAF and an optimal SAF for respectively medium and slow channel variations. We can notice that in both senarios, the BAF performs better than SAF. In the case of fast variations, optimal RLS and LMS shows very bad performance compared to slow variations.

In Fig. 6.6, we plot the BAF and SAF misadjustements vs. the SNR. Once again the BAF show better performance than the SAF.

As Fig 6.7 shows, the proposed Adaptive EM-Kalman algorithm converges to the optimal estimator. The convergence speed of the proposed algorithm is approximately the same as the of the conventional deterministic Kalman filtering (known parameters). Fig. 6.8 shows that the proposed Component-Wise adaptive Kalman algorithm converges in the steady-state to Kalman with known parameters. The convergence speed however is slower than in the cases of conventional adaptive EM-Kalman and Kalman algorithms. In the steady-state the two proposed algorithms outperform the existing stadard adaptive filtering approaches. Hence, the complexity of Component-Wise adaptive Kalman filter is linear in N, the adaptive filter order.

The good performances shown by the Component-Wise EM-Kalman motivates the study of next chapter where we invitigate algorithm with improved tracking and convergence propreties.

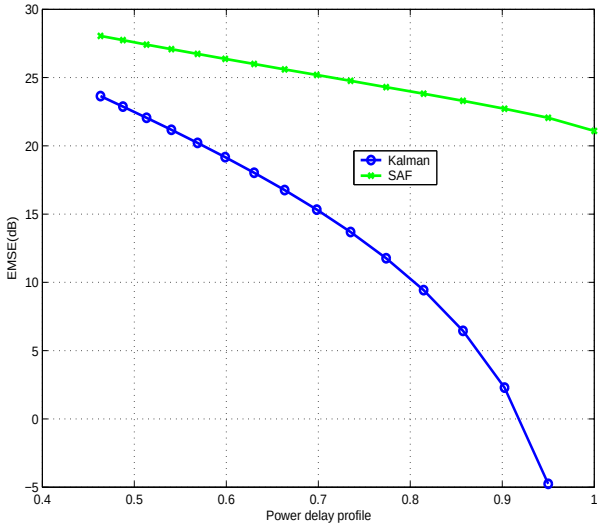

Figure 6.4: Comparative tracking performance results between BAF and Standard AF using EMSE for different value of power delay profile at $SNR = 15dB$ for a Medium variations ($A_i = 0.90$ $\tau = 0.5N$)

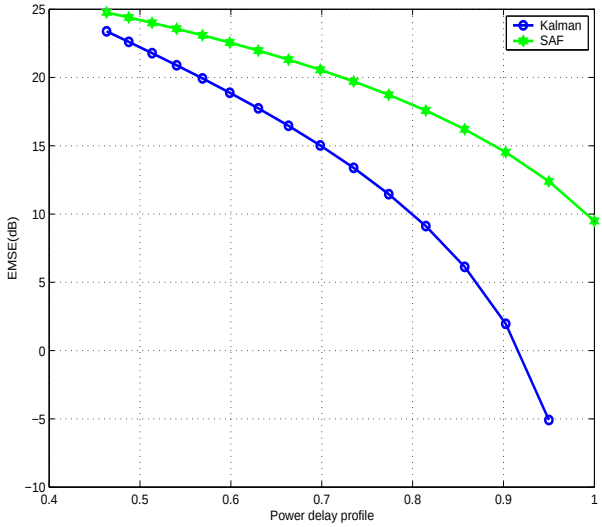

Figure 6.5: Comparative tracking performance results between BAF and Standard AF using EMSE for different value of power delay profile at $SNR = 15dB$ for a slow variations ($A_i = 0.90$ $\tau = 0.5N$)

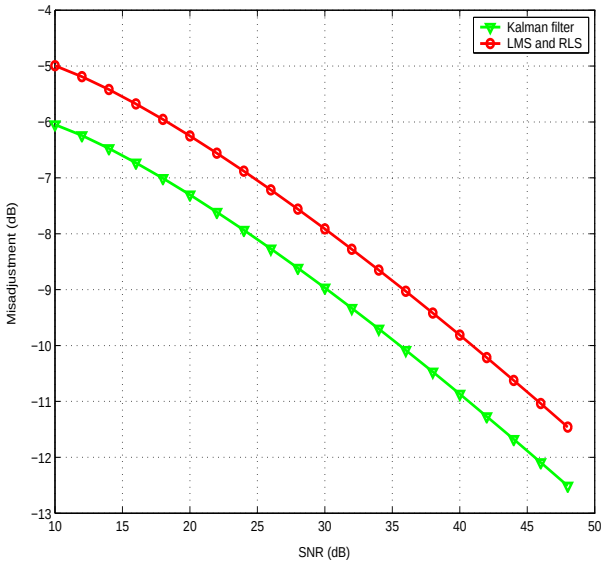

Figure 6.6: Comparison between the steady-state misadjustement of BAF and Standard AF

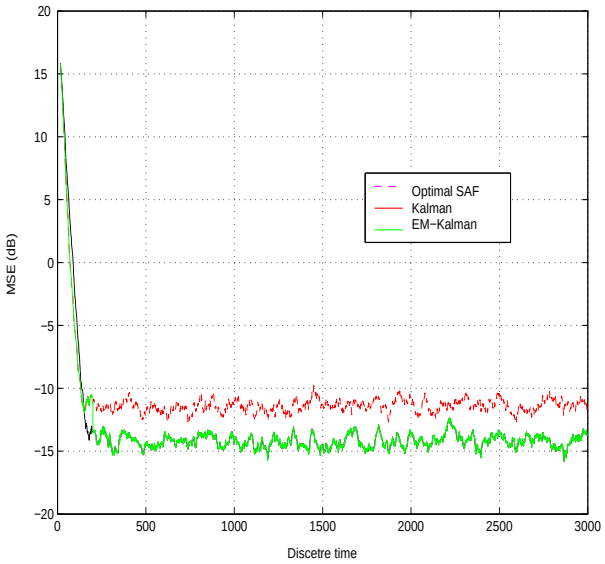

Figure 6.7: Comparison between the proposed Adaptive Kalman algorithm, SAF and Kalman filter SNR=20 dB, $\beta = 0.9$ and $\tau = 5N$

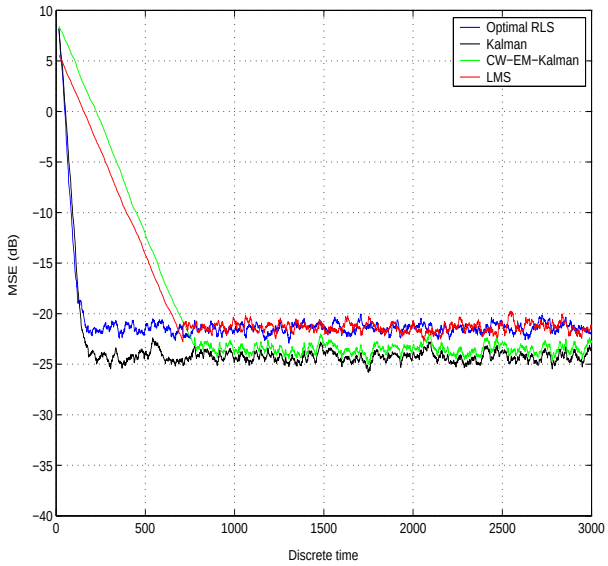

Figure 6.8: Comparison between the proposed CW-EM Adaptive Kalman algorithm and Kalman filter typical algorithms at SNR=15 dB, $\beta = 0.9$ and $\tau = 5N$

6.10 Conluding Remarks

In this chapter we studied the Bayesian Adaptive Filtering (BAF). We thus proposed different algorithms with incrementaly reduced complexity and with performances approaching thus of Kalman filter. We modeled the optimal adaptive filter coefficients variation as a stationary vector process, in particular as a AR(1) model. The filter parameters are adapted using the EM technique introduced in chapter 4 which has a complexity of the order of $O(N)$. We also derived analytical expressions for the EMSE for the different proposed algorithms. An evaluation in the radio mobile channel showed the good performances of our proposed techniques.
In the next chapter, we continue our study of BAF to propose improved algorithms with better tracking and convergence properties with reduced complexity.

Appendix

By using the convergence properties of the EM based algorithm developped in the chapter 4, the study the convergence of the EM technique used in the proposed CW EM-Kalman

algorithm.

The system 6.15 becomes $for\ n = 1 \dots N$, where N is the length of the filter

$$h_{k,n} = a_n h_{k-1,n} + w_{k,n} \tag{6.42}$$

$$y_k = h_{k-1,n} x_{k,n} + \sum_{j \neq n}^{N} h_{k-1,n} x_{k,n} + v_k \tag{6.43}$$

we can write

$$y_k - \sum_{j \neq n}^{N} \hat{h}_{k-1,n} x_{k,n} = h_{k-1,n} x_{k,n} + \sum_{j \neq n}^{N} \tilde{h}_{k-1,n} x_{k,n} + v_k$$

In each iteration y_k and v_k are updated as follows

$$y'_k = y_k - \sum_{j \neq n}^{N} \hat{h}_{k-1,n} x_{k,n}$$

and

$$v'_k = \sum_{j \neq n}^{N} \tilde{h}_{k-1,n} x_{k,n} + v_k$$

where w_k and v'_k are sequences of scalar-valued i.i.d. random variables distributed as $E[w_k w_k^T] = q$ and $E[v'_k v'^T_k] = r$.

for convenience, the parameters of this system shall be collected into the optimal vector $\theta^o = [a^o \ \ q^o]^T$

In order to avoid problems with the information matrices becoming unbounded as we allow the number of data to tend to infinity, we shall, entirely equivalently, employ the average value of this information matrix per sample is defined as

$$\begin{aligned} \bar{\Gamma}_{aug} &= \lim_{N \to \infty} \frac{1}{N} \Gamma_{aug} \\ &= \lim_{N \to \infty} \frac{-1}{N} [\frac{\partial^2}{\partial \theta \partial \theta^T} U(\theta, \theta^o)]|_{\theta = \theta^o} \\ &= \begin{pmatrix} \frac{1}{1-(a^o)^2} & 0 \\ 0 & \frac{q^o}{r(1-(a^o)^2)} \end{pmatrix} \end{aligned} \tag{6.44}$$

Discusion

The global rate of convergence of the EM algorithm is determined by the eigenvalue of Γ_{aug} small eigenvalues imply fast convergence. Since the eigenvalues of a diagonal matrix

are its diagonal elements it's quite clear, from equation (6.44), how the rate convergence of the algorithm is affected by the system parameters, as the number of data tends to infinity.

The first diagonal element of the matrix in equation (6.44) will be small if $a^o << 1$, that is, if the underlying system has fast dynamics.

Chapter 7

A Two Stage Approach to BAF

In the previous chapters, we proposed different Bayesian techniques with different complexities. Thus we proposed a EM-Kalman algorithm with complexity $O(N^2)$. To reduce the complexity, we presented the adaptive component-wise EM-Kalman technique with complexity $O(N)$ but which shows performance limitation in termes of tracking and convergence compared the previous technique. This motivated our study for the developpement of another approach with the same performance as the adaptive EM-Kalman but with the same complexity as the component-wise EM-Kalman.

The proposed two-stage algorithm consists of a first step employing a basic fast tracking adaptive filter, followed by lowpass filtering and downsampling of the time-varying filter coefficients. The second step then applies Kalman filtering at the reduced rate on a simplified state-space model, with an additive white noise measurement equation. The parameters in the state equation can be conveniently identified with an adaptive EM algorithm. The first stage would typically employ a (Normalized) LMS algorithm with a large stepsize. The main assumption underlying the proposed two-stage approach is that even in fast tracking applications, the bandwidth of the optimal filter variation is typically small compared to the signal bandwidth, motivating the downsampling operation. The first stage attempts to provide a bias-free filter estimate whereas the second stage optimizes the estimation variance. The performance of the proposed scheme is evaluated by simulations.

7.1 Introduction

Adaptive filtering is essentially intended for tracking time-varying optimal filters. The time variation of the optimal filter can be described by either expanding the filter coefficients into fixed time-varying (e.g. sinusoidal) basis functions (basis expansion models (BEMs)) [24] or by modeling them as stationary processes. The latter approach is perhaps better suited for minimum delay online processing. This case of constant slow variation of the filter coefficients (”drifting” parameters) is to be contrasted with another possible case of only occasional but significant variation (”jumping” parameters) which shall not be considered here. A lot of work has been done on optimizing the single parameter regulating the tracking speed of classical LMS or exponentially weighted RLS algorithms [1],[6]. For LMS, such an adaptive optimization leads to the class of Variable Step-Size (VSS) algorithms, see e.g. [65] and references therein. Adaptive filtering algorithms with a single adaptation parameter do not take into account that different portions of the filter may have different variation speeds and/or different magnitudes and hence are quite suboptimal. One noteworthy attempt to overcome this limitation is the introduction of a coefficient-wise VSS, as in [40], but the automatic adaptation of these VSSs is a difficult task.

In Bayesian Adaptive Filtering (BAF), prior information on the filter coefficient variances and variation spectra is exploited to optimize adaptive filter performance. A straightforward way to implement BAF is to use the Kalman filter, see e.g. [94],[13]. However, the complexity of the Kalman filter is enormous compared to that of the popular LMS adaptive filtering algorithm. Furthermore, the Kalman filter needs to be augmented with a state-space model identification technique.

Consider now the prototype adaptive filtering set-up, which is the system identification set-up, in which the desired-response signal d_k is modeled as the output of the optimal filter, which can be time-varying, plus independent (white) noise:

$$d_k = X_k^H H_k + v_k \tag{7.1}$$

where $X_k^H = [x_k \; x_{k-1} \cdots x_{k-N+1}]$ is the input signal vector and all terms are complex-valued. The input vector X_k is known up to time k and is assumed stationary with zero

mean and nonsingular covariance matrix $R = E[X_k X_k^H]$ Our aim is to estimate the time-varying parameter column vector H_k. Some general references on the tracking behavior of adaptive filtering algorithms are [1], [6],[18],. In this work we consider Bayesian Adaptive filtering based on a two-stage approach. A first stage with a fast standard adaptive filter, e.g. NLMS with stepsize equal to one. After some possible downsampling then, we consider an optimal filter in the second stage to extract H_k from the NLMS estimates, see figure (7.1).

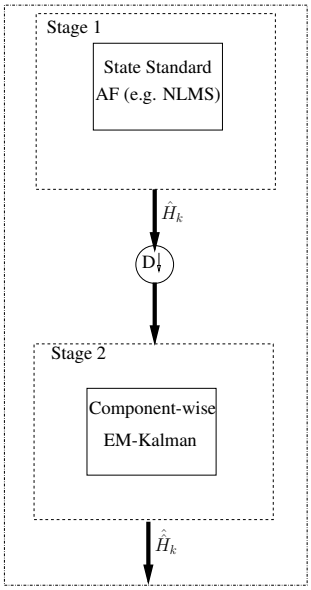

Figure 7.1: Two-stage adaptive filtering.

7.2 Stage 1: NLMS Algorithm

The simplest choice for a fast converging adaptive filtering (AFing) algorithm is a LMS algorithm with large stepsize, preferably the NLMS algorithm with normalized stepsize equal to one, or a smaller value of that order of magnitude. The NLMS algorithm updates the adaptive filter coefficients according to

$$e_k = d_k - X_k^H \hat{H}_{k-1} \tag{7.2}$$

$$\hat{H}_k = \hat{H}_{k-1} + \frac{\mu}{\|X_k\|^2} X_k e_k^* \tag{7.3}$$

For colored input and a FIR filter of length N, NLMS converges in general with N different modes that are of the form [10]

$$\frac{1}{1-\sqrt{1-\mu(2-\mu)\,\frac{\lambda_i}{\mathrm{tr}\,R}}\,q^{-1}} \tag{7.4}$$

where μ is the NLMS stepsize, we have assumed $N \gg 1$, the λ_i are the eigenvalues of the input signal covariance matrix $R = R_{XX}$, and q^{-1} is the one sample delay operator: $q^{-1}\,x_k = x_{k-1}$. We shall call the variation bandwidth of the optimal filter the Doppler bandwidth, which is the customary terminology when the adaptive filter represents a wireless channel response. We are going to assume that the Doppler bandwidth is much smaller than the signal bandwidth. For simplicity, let us assume that the input signal is not too colored so that the Doppler bandwidth can be smaller than the bandwidth $f_i = \frac{\mu(2-\mu)}{2\pi}\,\frac{\lambda_i}{\mathrm{tr}\,R}$ of each of the eigne modes. In this case, the NLMS adaptive filtering algorithm will pass the optimal filter coefficients undistortedly (zero bias). It will only introduce an estimation noise. In steady-state, this estimation noise leads to an estimation error $\widetilde{H}_k = H_k - \widehat{H}_k$ with covariance matrix $R_{\widetilde{H}\widetilde{H}} = \dfrac{\mu}{2-\mu}\,\dfrac{\sigma_v^2}{\mathrm{tr}\;R}\,I_N$. So the errors on the various filter components are uncorrelated and of identical variance. The errors are not temporally white however, due to the coloring introduced by the filtering of the modes in (7.4). However, due to the previous assumptions, the estimation noise can be considered white over the Doppler bandwidth of the optimal filter.

A better alternative to the NLMS algorithm in the first stage would be an adaptive filter that is less sensitive to the input signal color. If we want no such sensitivity then a Recursive Least-Squares (RLS) algorithm should be used. To minimize the distortion (so-called "lag noise") on the optimal filter the best RLS choice would be one with a sliding rectangular window, in which a delay gets introduced equal to half the window length (non-causal adaptive filtering) [147]. RLS algorithms are more complex that (N)LMS, but fast versions exist. There is also a whole range of adaptive filtering algorithms in between LMS and RLS in terms of complexity and performance, such Affine Projection Algorithms, Fast Newton Transversal Filters, frequency domain adaptive filters, LMS with prewhitening etc.

7.3 Subsampling Glue

As mentioned earlier, if the Doppler bandwidth is significantly less than the signal bandwidth (sampling rate), then it would be overkill to put in place an optimal tracking algorithm working at the sampling rate. In that case, the output of the first stage (the vector sequence $\widehat{H}_k$) can be lowpass filtered and commensurate downsampled without introducing distortion (lag noise) as long as the lowpass filter does not distort the Doppler spectrum. The main goal of this operation is to reduce complexity. Indeed further processing in the second stage can now be performed at a reduced rate. And fixed lowpass filtering does not have to be a complex operation (if a simple filter is used, for instance first order IIR (exponential averaging)). Another reason is that, whereas it would constitute quite an approximation to model $\widetilde{H}_k$ as temporally white, after lowpass filtering and downsampling (with a factor D), such an approximation becomes more accurate. The lowpass filtering operation reduces the estimation noise roughly with a factor D. In what follows, we shall continue to use the same notation for the subsampled rate and continue to denote the lowpass filtered and subsampled NLMS output as $\widehat{H}_k$. This provides the measurement data for stage two.

7.4 Stage 2: "Diagonal" EM-Kalman Filtering

Consider the state-space model

$$H_{k+1} = AH_k + W_k \tag{7.5}$$

$$\hat{H}_k = H_k + \tilde{H}_k \tag{7.6}$$

The measurement and process noise terms are assumed to be zero mean Gaussian with covariances $R_{\tilde{H}}$ and Q respectively.The matrix A contains information about how the states evolve. It is particularly useful in tracking applications. The matrix A should be viewed as a mechanism to achieve directed trajectories in state space. In other words, A allows for more general jumps than the simple random walk that would result by excluding A from the model. Despite the fact that the data is processed in batches, the model of equation (7.6) allows the weights to be time varying. It is, therefore, possible to deal with

non-stationary data sets. In the event of the data being stationary, we should expect the process noise term to vanish. Consequently, if we know that the data is stationary, the estimate of the process noise can be used to determine how well the model explains the data.

The objective is to estimate the model states (weights)H_k and the set of parameters $\phi = \{A, Q, R\}$ given the measurements $\hat{H}_{1:N}$. Then we use a Kalman smoother to estimate H_k and EM algorithm to estimate the set of parameters. Since the Kalman model is diagonal, we propose a Component-Wise Adaptive Kalman algorithm to update the filter coefficients, which decreases computational complexity. Then the model (7.6) becomes:

$$h_{i+1} = a\, h_i + w_i \tag{7.7}$$

$$\hat{h}_i = h_i + \tilde{h}_i \tag{7.8}$$

7.4.1 Kalman smoother

Smoothing often entails forward and backward filtering over a segment of data so as to obtain improved averaged estimates. Various techniques have been proposed to accomplish this goal . This study uses the well-known Rauch-Tung-Striebel smoother . The forward filtering stage involves computing the estimates $\hat{h}_k$ and P_k , over a segment of I samples, with the following KF recursions:

$$\hat{\hat{h}}_{i+1|i} = a\hat{\hat{h}}_i \tag{7.9}$$

$$p_{i+1|i} = aa^* p_i + q \tag{7.10}$$

$$k^f_{i+1} = p_{i+1|i}(r + p_{i+1|i})^{-1} \tag{7.11}$$

$$\hat{\hat{h}}_{i+1} = \hat{\hat{h}}_{i+1|i} + k^f_{i+1}(\hat{h}_{i+1} - \hat{\hat{h}}_{i+1|i}) \tag{7.12}$$

where k^f denotes the Kalman gain . Subsequently, the Rauch- Tung-Striebel smoother makes use of the following backward recursions:

$$J_{i-1} = \frac{p_{i-1}a^*}{p_{i|i-1}} \tag{7.13}$$

$$\hat{\hat{h}}_{i-1|n} = \hat{\hat{h}}_{i-1} J_{i-1}(\hat{\hat{h}}_{i|n} - a\hat{\hat{h}}_{i-1}) \tag{7.14}$$

$$p_{i-1|n} = p_{i-1} + J_{i-1}(p_{i|n} - p_{i|i-1})J^*_{i-1} \tag{7.15}$$

$$p_{i,i-1|n} = p_i J^*_{i-1} + J_i(p_{i+i|n} - ap_i)J^*_{i-1} \tag{7.16}$$

where the parameters, covariance and cross-covariance are defined as follows:

$$\begin{aligned} \hat{\hat{h}}_{i+1|n} &= E[h_{i+1} \mid \hat{h}_{1:n}] \\ p_{i|n} &= E[(h_i - \hat{\hat{h}}_i)(h_i - \hat{\hat{h}}_i)^* \mid \hat{h}_{1:n}] \\ p_{i,i-1|n} &= E[(h_i - \hat{\hat{h}}_i)(h_{i-1} - \hat{\hat{h}}_{i-1})^* \mid \hat{h}_{1:n}] \end{aligned} \tag{7.17}$$

7.4.2 Model parameters adaptation

The state model parameters can be adapted using the EM algorithm introduced in chapter 4, according to

$$\begin{aligned} \psi_{i|i} &= \lambda\psi_{i,n|i-1} + (\hat{\hat{h}}_{i|i}\hat{\hat{h}}^*_{i|i} + p_{i|i}) \\ \psi_{i-1|i} &= \lambda\psi_{i-1|i-1} + (\hat{\hat{h}}_{i-1|i}\hat{\hat{h}}^*_{i-1|i} + p_{i-1|i}) \\ d_i &= p_{i|i}C^*_{i-1} \\ &= a_i p_{i-1|i-1} - k^f_i(a_i^{-1}(1 - q_i x_i))^* \end{aligned} \tag{7.18}$$

$$\begin{aligned} \psi_{i,i-1|i} &= \lambda\psi_{i,i-1|i-1} + (\hat{\hat{h}}_{i|i}\hat{\hat{h}}^*_{i-1|i} + d_i) \\ q_{i+1} &= \frac{1}{\gamma_i}(\psi_{i,|i} - \frac{\psi_{i,i-1|i}}{(\psi_{i-1|i})}(\psi_{i,i-1|i})^* \\ a_{i+1} &= \psi_{i,i-1|i}(\psi_{i-1|i})^{-1} \end{aligned} \tag{7.19}$$

7.4.3 Steady-State Excess Mean-Square Error (EMSE)

The state estimate update is given by Kalman as :

$$\begin{aligned} \hat{\hat{H}}_{k|k} &= A\hat{\hat{H}}_{k-1|k-1} + K_k(\hat{H}_k - A\hat{\hat{H}}_{k-1|k-1}) \\ &= A\hat{\hat{H}}_{k-1|k-1} + K_k(H_{k-1} - A\hat{\hat{H}}_{k-1|k-1}) + K_k e_{opt} \\ &= A\hat{\hat{H}}_{k-1|k-1} + \frac{P_{k|k}}{\sigma_v^2}(H_{k-1} - A\hat{\hat{H}}_{k-1|k-1}) \\ &\quad + K_k e_{opt} \end{aligned} \tag{7.20}$$

Where $K_k = P_{k|k} R^{-1}_{\tilde{H}_k \tilde{H}_k}$
and e_{opt} represents the minimum(in a mean square sense) error at time k. In studying tracking behavior, we may exclude the influence of the estimation noise, since the deviation of $E[H_{k|k}]$ from H_k determines the response of the BAF algorithm to the non-stationarity of the environment. Taking expected values on both sides of (7.20), we get

$$\begin{aligned} E[\hat{\hat{H}}_{k|k}] &= AE[\hat{\hat{H}}_{k-1|k-1}] \\ &\quad + P_{k|k} R^{-1}_{\tilde{H}_k \tilde{H}_k} (H_{k-1} - AE[\hat{\hat{H}}_{k-1|k-1}]) \\ &= AE[\hat{\hat{H}}_{k-1|k-1}] \\ &\quad + P_{k|k} R^{-1}_{\tilde{H}_k \tilde{H}_k} (H_{k-1} - AE[\hat{\hat{H}}_{k-1|k-1}]) \end{aligned} \tag{7.21}$$

the lag-error is given by

$$\tilde{L}_k = E[\hat{\hat{H}}_{k|k}] - H_k$$

$$\begin{aligned} \tilde{L}_k &= AE[\hat{\hat{H}}_{k-1|k-1}] - P_{k|k} A R^{-1}_{\tilde{H}_k \tilde{H}_k} (\tilde{H}_{k-1}) \\ &\quad + P_{k|k} R^{-1}_{\tilde{H}_k \tilde{H}_k} (H_{k-1} - AH_{k-1}) - H_k + AH_{k-1} \\ &= AE[\hat{\hat{H}}_{k-1|k-1}] - P_{k|k} A R^{-1}_{\tilde{H}_k \tilde{H}_k} (\tilde{H}_{k-1}) \\ &\quad + (\frac{P_{k|k} R}{\sigma_v^2} (I - A) + A) H_{k-1} - H_k \\ &\approx (I - P_{k|k} R^{-1}_{\tilde{H}_k \tilde{H}_k}) A \tilde{H}_{k-1} + AH_{k-1} - H_k \end{aligned} \tag{7.22}$$

The input is considered to be white ($R = \sigma_x^2 I$), note that, each element of the lag-error vector is determined by the following relation:

$$\tilde{l}_i(k) = (1 - \frac{p_{i,k|k} N \sigma_x^2}{D \sigma_v^2}) A_i \tilde{l}_{i,k-1} + A_i h_{i,k-1} - h_{i,k}. \tag{7.23}$$

where $\tilde{l}_i(k)$ is the i^{th} element of $\tilde{L}_k$. By properly interpreting the equation above, we can say that the lag is generated by applying the transformed instantaneous optimal coefficient to a first-order discrete-time filter denoted ***lag filter***

$$\tilde{L}_i(z) = \frac{A_i z^{-1} - 1}{1 - (1 - \frac{N \sigma_x^2 p_{i,k|k}}{D \sigma_v^2}) A_i z^{-1}} H_i(z) \tag{7.24}$$

Using the inverse z-transform, the variance of the elements of the vector $\tilde{L}(k)$ can then be calculated by

$$E[\tilde{l}_i(k) \tilde{l}_i^H(k)] = \frac{1}{2\pi j} \oint \tilde{L}_i(z) \tilde{L}_i(z^{-1}) q_i z^{-1} dz$$

The BAF excess mean square error due to lag is then given by Eq. (7.25)

$$EMSE = \sum_{i=1}^{N} EMSE_i \tag{7.25}$$

where

$$EMSE_i = \sigma_x^2 p_i = \sigma_x^2 \frac{-(\frac{Dq_i}{NSNR_i} + q_i) + q_i\sqrt{(1 + \frac{D}{NSNR_i})^2 + 4\,\frac{N\sigma_x^2}{D\sigma_v^2} \mid a_i \mid^2}}{2 \mid a_i \mid^2} \tag{7.26}$$

7.5 Conclusion

The choosen application is again the mobile radio channel as in the previous chapter. The behavior of two-stage adaptive, NLMS and Kalman algorithms are compared on the basis of simulation results, as shown in Fig. 7.2. As this figure shows, the proposed two-stage adaptive algorithm converges to Kalman filter (known parameters). It offers better performance than NLMS algorithm in steady-state.

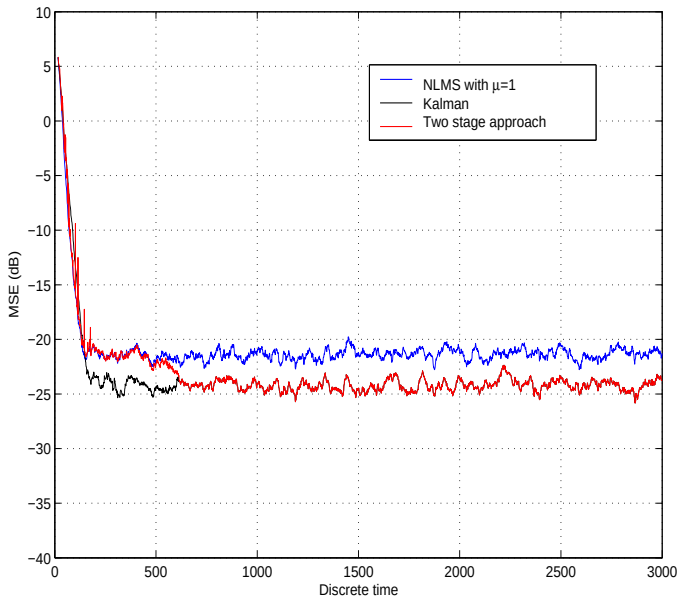

Figure 7.2: Comparison between the proposed two-stage adaptive filter, NLMS and the Kalman filter with known optimal parameters.

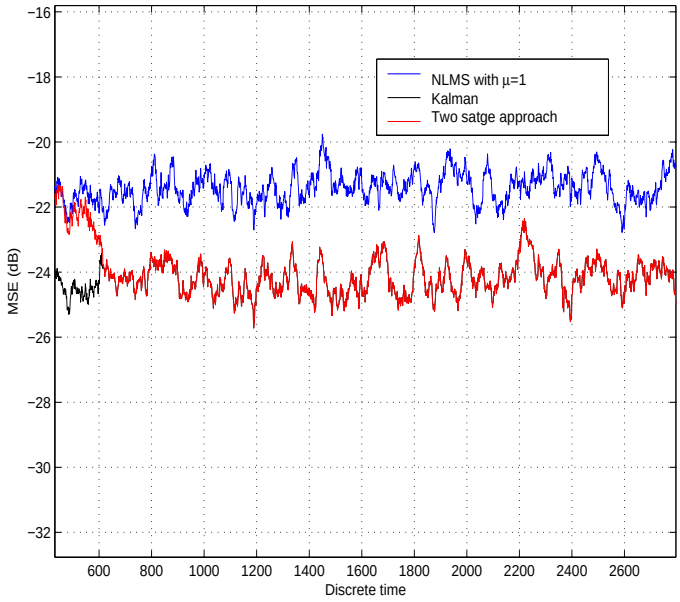

Figure 7.3: Zoom on the steady-state behavior.

The algorithm is presented in the following Table of component-wise Adaptive EM-

Kalman.

Two Stage Algorithm	
Computation	**Cost** $(\times)$
Initialization	
$\hat{\hat{h}}_{0\|0} = 0\ ,\ \ P_{0\|0} = 100\ ,$	
$a_0 = \alpha\ ,\ q_0 = (1-\alpha)$	
$\pi_{0\|0} = 0\ ,\ \ \pi_{1,0\|0} = 0,\ \ \pi_{1\|0} = 0$	
$\gamma_0 = 0$	
NLMS	
$e_k = d_k - X_k^H\ \hat{H}_{k-1},$	
$\hat{H}_k = \hat{H}_{k-1} + \frac{\mu}{\|X_k\|^2}\ X_k\ e_k^*,$	
CW Kalman filtering and one step smoothing	
$for\ n = 1 \dots N$	
$\hat{\hat{h}}_{k,n\|k-1} = a_{n,k}\hat{\hat{h}}_{k-1,n\|k-1}$	1
$\hat{y}_{k,n\|k-1} = \hat{\hat{h}}_{k,n\|k-1}$	0
$K_{n,k} = P_{k,n\|k-1}$	0
$M_{k,n} = (K_{n,k} + \frac{N\sigma_x^2}{D\sigma_v^2})^{-1}$	1
$K_{n,k}^f = K_{n,k}M_{n,k}$	1
$C_{n,k-1} = P_{n,k-1\|k-1}a_{n,k}^* P_{n,k\|k-1}^{-1}$	2
$\hat{\hat{h}}_{n,k-1\|k} = \hat{\hat{h}}_{n,k\|k-1} + K_{n,k}^f(\hat{h}_k - \hat{y}_{k,n\|k-1})$	1
$P_{k,n\|k-1} = \mid a_{n,k} \mid^2 P_{k-1,n\|k-1} + q_{n,k}$	1
$\hat{\hat{h}}_{k,n\|k} = \hat{\hat{h}}_{k,n\|k-1} + a_{n,k}(K_k - q_{n,k})M_k(\hat{h}_k - \hat{y}_{k,n\|k-1})$	2
$P_{k,n\|k} = P_{k,n\|k-1} - K_{n,k}^f K_{n,k}^*$	1
$P_{k-1,n\|k} = P_{k-1,n\|k-1} + C_{n,k-1}(P_{n,k\|k} - P_{n,k\|k-1})$	1
Model Parameters Adaptation	
$\pi_{n,k\|k} = \lambda\pi_{k,n\|k-1} + \hat{\hat{h}}_{n,k\|k}\hat{\hat{h}}_{n,k\|k}^* + P_{n,k\|k}$	1
$\Pi_{n,k-1\|k} = \lambda\pi_{n,k-1\|k-1} + \hat{\hat{h}}_{n,k-1\|k}\hat{\hat{h}}_{n,k-1\|k}^H + P_{n,k-1\|k}$	1
$D_{n,k} = P_{n,k\|k}C_{n,k-1}^* = a_{n,k}P_{n,k-1\|k-1} - K_{n,k}^f(a_{n,k}^{-1}(K_{n,k} - q_{n,k}x_{n,k}))^*$	2
$\pi_{n,k,k-1\|k} = \lambda\pi_{n,k,k-1\|k-1} + diag(\hat{\hat{h}}_{n,k\|k}\hat{\hat{h}}_{n,k-1\|k}^* + D_{n,k})$	1
$q_{n,k+1} = \frac{1}{\gamma_k}(\pi_{n,k,\|k} - \pi_{n,k,k-1\|k}(\pi_{n,k-1\|k})^{-1}(\pi_{n,k,k-1\|k})^*)$	2
$\gamma_k = \gamma_{k-1} + 1$	
$a_{n,k+1} = \pi_{n,k,k-1\|k}(\pi_{n,k-1\|k})^{-1}$	1
cost/update $19N$	

APPENDIX

The fact that the trace and expectation operators are linear, the expectation of the log-likelihood becomes:

$$
\begin{aligned}
E[l(\phi))] \;=\; & -\frac{1}{2}\sum_{i=1}^{M} tr(q^{-1}[\hat{\hat{h}}_{i|M}\hat{\hat{h}}_{i|M}^{H} + p_{i|M} \\
& -2a(\hat{\hat{h}}_{i|M}\hat{\hat{h}}_{i-1|M}^{H} + p_{i,i-1|M})^{H} \\
& +a(\hat{\hat{h}}_{i-1|M}\hat{\hat{h}}_{i-1|M}^{H} + p_{i-1|M})a^{H}]) \\
& -\sum_{i=1}^{M}\frac{1}{2}tr(r^{-1}[\hat{h}_i\hat{h}_i^{H} \\
& -\hat{\hat{h}}_{i|M}\hat{h}_i^{H} - \hat{h}_i\hat{\hat{h}}_{i|M}^{H} \\
& \hat{\hat{h}}_{i|M}\hat{\hat{h}}_{i|M}^{H} + p_{i|M}]) - \frac{M}{2}\,|q| - \frac{M}{2}\,|r|
\end{aligned}
$$

So far, it has been shown that given a set of parameters $\phi = \{a, q, r\}$ and a measurements $\hat{h}$, it is possible to compute the expected values of the states with an Kalman smoother. This section presents an EM algorithm to learn the parameters ϕ.

The EM algorithm is an iterative method for finding a mode of the likelihood function $p(\hat{h}_{|}\phi)$. It proceeds as follows: (E-step 1) estimate the states h given a set of parameters ϕ, (M-step 1) estimate the parameters given the new states, (E-step 2) re-estimate the states with the new parameters, and so forth. The most remarkable attribute of the EM algorithm is that it ensures an increase in the likelihood function at each iteration. It is hard to maximize $p(\hat{h}_{|}\phi)$, EM will allow us to accomplish this by working with $p(\hat{h}_{,}h_{|}\phi)$. To gain more insight into the EM method, let us express the likelihood function as follows:

$$
\begin{aligned}
p(\hat{h} \mid \phi) \;&=\; p(\hat{h} \mid \phi)\frac{p(h \mid \hat{h}, \phi)}{p(h \mid \hat{h}, \phi)} && (7.27)\\
&=\; \frac{p(h, \hat{h} \mid \phi)}{p(h \mid \hat{h}, \phi)} && (7.28)\\
& && (7.29)
\end{aligned}
$$

Taking the logarithns of both sides yields the following identity:

$$
\begin{aligned}
\ln(p(\hat{h} \mid \phi)) \;=\; & \ln(p(h, \hat{h} \mid \phi)) \\
& -\ln(p(h \mid \hat{h}, \phi)).
\end{aligned}
$$

Let us treat h as a randon variable with distribution $p(h \mid \hat{h}, \phi^{old})$, where ϕ^{old}, is the current guess. If we then take expectations on both sides of the previous identity, while

renenbering that the left hand side does not depend on h, we get:

$$\begin{aligned} E[\ln(p(\hat{h} \mid \phi))] &= E[\ln(p(h, \hat{h} \mid \phi))] \\ &\quad -E[\ln(p(h \mid \hat{h}, \phi))] \end{aligned}$$

we need to develop an expression for the likelihood of the conpleted data. The likelihood of the data given the states, the initial conditions and the evolution of the states nay be approxinated by Gaussian distributions.

$$\begin{aligned} p(h_i \mid h_{i-1}, \phi) &= \frac{1}{(2\pi)^{\frac{n}{2}} |q|^{\frac{1}{2}}} \exp q^{-1}(\frac{1}{2}(h_i - ah_{i-1}) \\ &\quad \times(h_i - ah_{i-1})^*) \\ p(\hat{h}_i \mid h_i, \phi) &= \frac{1}{(2\pi)^{\frac{n}{2}} |r|^{\frac{1}{2}}} \exp(\frac{1}{2}(\hat{h}_i - \hat{\hat{h}}_i) \\ &\quad \times r^{-1}(\hat{h}_i - \hat{\hat{h}}_i)^*) \end{aligned} \tag{7.30}$$

Under the nodel assunptions of uncorrelated noise sources and narkov state evolution, the likelihood of the conplete data is given by:

$$\begin{aligned} p(h_{1:n}, \hat{h}_{1:n} \mid \phi) &= \prod_{i=1}^{n} p(h_i \mid h_{i-1}, \phi) \\ &\quad \times \prod_{i=1}^{n} p(\hat{h}_i \mid h_i, \phi) \end{aligned}$$

Hence, the log-likelihood of the complete data is given by the following expression:

$$E[\ln(p(h_{1:n}, \hat{h}_{1:n} \mid \phi))] = l(\phi)$$

$$\begin{aligned} l(\phi) &= -\sum_{k=1}^{n} q^{-1} \frac{1}{2}(h_i - ah_{i-1})(h_i - ah_{i-1})^* \\ &\quad -\sum_{k=1}^{n} r^{-1} \frac{1}{2}(\hat{h}_i - \hat{\hat{h}}_i)(\hat{h}_i - \hat{\hat{h}}_i)^* \\ &\quad -\frac{n}{2}|q| - \frac{n}{2}|r| \end{aligned} \tag{7.31}$$

As discussed in the previous section, all we need to do now is to compute the expectation of $\ln(p(h_{1:n}, \hat{h}_{1:n} \mid \phi))$ and then differentiate the result with respect to the parameters ϕ so as to maximize it. The EM algorithm for nonlinear state space models will thus involve computing the expected values of the states and covariances with the Kalman smoother and then maximizing the parameters ϕ with the formulae obtained by differentiating the expected log-likelihood.

7.5.1 Computing the expectation of the log-likelihood

The derivation requires the following sufficient statistics:

$$E[h_i \mid \hat{h}_{1:n}] = \hat{\hat{h}}_{i|n} \tag{7.32}$$

$$E[h_i h_i \mid \hat{h}_{1:n}] = p_{i|n} + \hat{\hat{h}}_{i|n}\hat{\hat{h}}^*_{i|n} \tag{7.33}$$

$$E[h_i h_{i-1} \mid \hat{h}_{1:n}] = p_{i,i-1|n} + \hat{\hat{h}}_{i|n}\hat{\hat{h}}^*_{i-1|n} \tag{7.34}$$

Now, taking the expectation of the log-likelihood for the complete data, by averaging over $h_{1:n}$ under the distribution $p(h_{1:n} \mid, \hat{h}_{1:n}, \phi^{old})$, one gets the following expression:

$$\begin{aligned} l(\phi) = & -\frac{1}{2q}\sum_{i=1}^{n} E[h_i^* h_i - h_i^* a h_{i-1} \\ & -(a^* h_{i-1})^* h_i \\ & +(a^* a h_{i-1})^* h_{i-1}] \\ & -\sum_{i=1}^{n} \frac{1}{2r} E[\hat{h}_i^* r^{-1} \hat{h}_i - \hat{h}_i^* \hat{\hat{h}}_{i|i-1} \\ & -\hat{\hat{h}}^*_{i|i-1}\hat{h}_i + \hat{\hat{h}}^*_{i|i-1}\hat{\hat{h}}_{i|i-1}] \\ & -\frac{n}{2}|q| - \frac{n}{2}|r| \end{aligned}$$

Completing squares and using the following abbreviations:

$$\pi_{i|n} = \sum_{i=1}^{n} \hat{\hat{h}}_{i|n}\hat{\hat{h}}^*_{i|n} + p_{i|n} \tag{7.35}$$

$$\pi_{i-1|n} = \sum_{i=1}^{n} \hat{\hat{h}}_{i-1|n}\hat{\hat{h}}^*_{i-1|n} + p_{i-1|n} \tag{7.36}$$

$$\pi_{i,i-1|n} = \sum_{i=1}^{n} \hat{\hat{h}}_{i|n}\hat{\hat{h}}^*_{i-1|n} + p_{i,i-1|n} \tag{7.37}$$

7.5.2 Differentiating the expected log-likelihood

Maximum with respect to a: Differentiating the expected log-likelihood with respect to a yields:

$$\frac{\partial E[l(\phi)]}{\partial a} = \frac{-1}{2q}(-2\Pi_{i,i-1|n} + 2a\Pi_{i-1|n}) \tag{7.38}$$

Equating this result to zero yields the value of A that maximizes the approximate log-likelihood:

$$a = \Pi_{i,i-1|n}(\Pi_{i-1|i})^{-1} \tag{7.39}$$

Maximum with respect to $r = r_{\tilde{H}}$: Differentiating the expected log-likelihood with respect to r^{-1} gives:

$$\begin{aligned}\frac{\partial E[l(\phi)]}{\partial r^{-1}} &= -\sum_{i=1}^{n}\frac{1}{2}((\hat{h}_i - \hat{\hat{h}}_{i|n})(\hat{h}_i - \hat{\hat{h}}_{i|n})^* \\ &\quad + p_{i|n}) + \frac{n}{2}r\end{aligned} \tag{7.40}$$

Hence, by equating the above result to zero, the maximum of the approximate log-likelihood with respect to r is given by:

$$r = \sum_{i=1}^{n}\frac{1}{n}((\hat{h}_i - \hat{\hat{h}}_{i|n})(\hat{h}_i - \hat{\hat{h}}_{i|n})^* + p_{i|n}) \tag{7.41}$$

Maximum with respect to q: Maximum with respect to q Following the same steps, the derivative of the expected log-likelihood with respect to q^{-1} is given by:

$$\begin{aligned}\frac{\partial E[l(\phi)]}{\partial r^{-1}} &= -\frac{1}{2}(\pi_{i|n} - 2a\pi^*_{i,i-1|n} + a\pi_{i-1|n}a^*) \\ &\quad + \frac{n}{2}q\end{aligned}$$

Hence, equating to zero and using the result that
$a = \frac{\pi_{k,k-1|n}}{(\Pi_{k-1|n})}$, the maximum of the approximate log-likelihood with respect to q is given by:

$$q = \frac{1}{n}(\pi_{i|n} - \pi_{i,i-1|n}\pi_{i-1|n}\pi^*_{i,i-1|n}) \tag{7.42}$$

Part II

Window Optimization Issues in Recursive Least-Squares Adaptive Filtering And Tracking

Chapter 8

Window Optimization Issues in Recursive Least-Squares Adaptive Filtering And Tracking

8.1 Introduction

In this chapter we consider tracking of an optimal filter modeled as a stationary vector process. We interpret the Recursive Least-Squares (RLS) adaptive filtering algorithm as a filtering operation on the optimal filter process and the intantaneous gradient noise (induced by the measurement noise). The filtering operation carried out by the RLS algorithm depends on the window used in the least-squares criterion. To arrive at a recursive LS algorithm requires that the window impulse response can be expressed recursively (output of an IIR filter). In practice, only two popular window choices exist (with each one tuning parameter): the exponential weighting (W-RLS) and the rectangular window (SWC-RLS). However, the rectangular window can be generalized at a small cost for the resulting RLS algorithm to a window with three parameters (GSW-RLS) instead of just one, encompassing both SWC- and W-RLS as special cases. Since the complexity of SWC-RLS essentially doubles with respect to W-RLS, it is generally believed that this increase in complexity allows for some improvement in tracking performance. We show that, with

equal estimation noise, W-RLS generally outperforms SWC-RLS in causal tracking, with GSW-RLS still performing better, whereas for non-causal tracking SWC-RLS is by far the best (with GSW-RLS not being able to improve). When the window parameters are optimized for causal tracking MSE, GSW-RLS outperforms W-RLS which outperforms SWC-RLS. We also derive the optimal window shapes for causal and non-causal tracking of arbitrary variation spectra. It turns outs that W-RLS is optimal for causal tracking of AR(1) parameter variations whereas SWC-RLS if optimal for non-causal tracking of integrated white jumping parameters, all optimal filter parameters having proportional variation spectra in both cases.

The RLS algorithm is one of the basic tools for adaptive filtering. The convergence behavior of the RLS algorithm is now well understood. Typically, the RLS algorithm has a fast convergence rate, and is not sensitive to the eigenvalue spread of the correlation matrix of the input signal. However, when operating in a non-stationary environment, the adaptive filter has the additional task of tracking the variation in environmental conditions. In this context, it has been established that adaptive algorithms that exhibit good convergence properties in stationary environments do not necessarily provide good tracking performance in a non-stationary environment; because the convergence behavior of an adaptive filter is a transient phenomenon, whereas the tracking behavior is a steady-state property [137, 138].

One fundamental non-stationary scenario involves a time-varying system in which the cross-correlation between the input signal and the desired response is time-varying. This case occurs in the system identification setup. To take into account system variation, two main variants of RLS algorithms exist. The first introduces a forgetting factor, and leads to the exponentially Weighted RLS (W-RLS) approach. The second uses a Sliding rectangular Window (SWC-RLS approach). In [139, 140], a generalized sliding window RLS (GSW-RLS) algorithm was introduced, that generalizes the W-RLS and SWC-RLS algorithms. The GSW-RLS uses a generalized window (see Fig. 8.1), which consists of an exponential window with a discontinuity at delay L. It can be seen that the exponential and rectangular windows are particular cases of the generalized window, for $\alpha = 0$ and $(\alpha, \lambda) = (1, 1)$ resp.

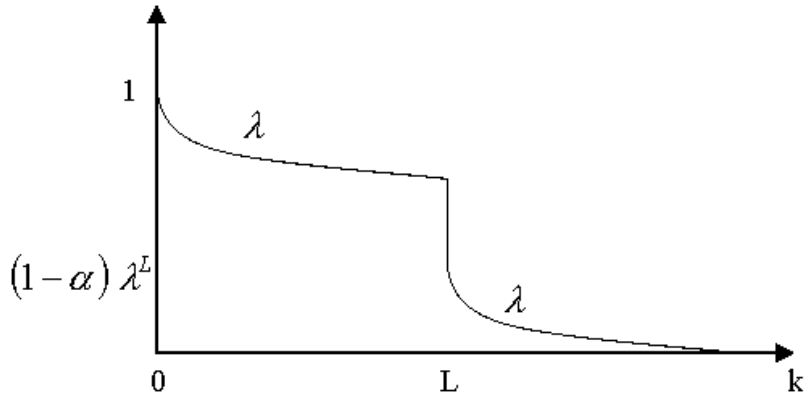

Figure 8.1: The generalized sliding window

In [139, 140], a tracking improvement for GSW-RLS was observed for different system variation models (AR(1), MA, and Random walk). On the one hand the initial portion of the window permits to emphasize the very recent past which allows very fast tracking. On the other hand, the GSW-RLS algorithm solves nevertheless an overdetermined system of equations and hence enjoys the fast convergence properties of RLS algorithms. Another effect of the exponential tail of the GSW is regularization. In fact, the rectangular window sample covariance matrix appearing in SWC-RLS can be particularly ill-conditioned compared to a sample covariance matrix based on an exponential window with compatible time constant. Finally, the GSW-RLS algorithm turns out to have the same structure and comparable computational complexity as the SWC-RLS algorithm.

This chapter is organized as follows. In section 8.2, a tracking analysis in the frequency domain is presented. Uninformed and Informed Bayesian approaches are investigated respectively in sections 8.3, and 8.4. Finally a discussion and concluding remarks are provided in section 8.5.

8.2 Tracking Characteristics of RLS Algorithms

We consider the classic adaptive system identification problem (see Fig. **??**). The adaptive system identification is designed for determining a (typically linear FIR) model of the transfer function for an unknown, time-varying digital or analog system.

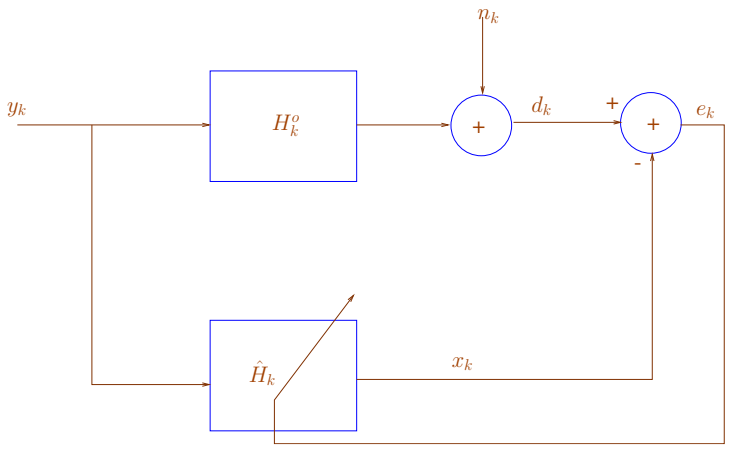

Figure 8.2: System identification block diagram

The adaptive system identification problem can be described by:

$$\begin{cases} d_k = H_k^{oT} Y_k \ + \ n_k \\ x_k = H_k^T Y_k \end{cases} \tag{8.1}$$

where

- n_k is an iid Gaussian noise sequence $(n_k \sim \mathcal{N}(0, \sigma_n^2))$ where σ_n^2 is the Minimum Mean Squared Error (MMSE)
- H_k^o denotes the optimal Wiener Filter
- H_k represents the adaptive Filter
- e_k is the a posteriori error given by:

$$e_k = d_k - x_k = \widetilde{H}_k^T Y_k \ + \ n_k \tag{8.2}$$

where $\widetilde{H}_k = H_k^o - H_k$ denotes the filter deviation.

In weighted RLS, the set of the N adaptive filter coefficients $H_k = [H_{1,k} \cdots H_{N,k}]^T$

gets adapted so as to minimize recursively the Weighted Least Squares criterion

$$J_k = F(q)\, e_k^2 = \sum_i f_i\, e_{k-i}^2 \tag{8.3}$$

where $F(z) = \sum_i f_i\, z^{-i}$ is the transfer function of the weighting window f_i characterizing the RLS algorithm, and $q^{-1}\, e_k^2 = e_{k-1}^2$. There are a number of references dealing with the performance of RLS algorithms in non-stationary environments [143, 142, 141, 140]. The basic idea is to focus on the model quality in terms of the output Excess MSE (EMSE). We consider stationary optimal filter variation models, hence the RLS algorithm will reach a stationary regime to which we limit attention. The EMSE is defined as:

$$EMSE \;=\; E\left\{e_k^2\right\} - \sigma_n^2 \;=\; E\left\{Y_k^T \widetilde{H}_k \widetilde{H}_k^T Y_k\right\} \tag{8.4}$$

(in principle the a priori error signal should be considered for the EMSE, we shall stick to the a posteriori error signal to avoid the appearance of a delay in the notation). So, if we assume that the system variation is a zero-mean, wide-sense stationary process H_k^o with a power spectral density matrix $S_{HH}\left(e^{j2\pi f}\right)$, and if we invoke the independence assumption, in which Y_k and $\widetilde{H}_k$ are assumed to be independent (this works better for the a priori error), the EMSE can be expressed in the following form:

$$EMSE = \; tr\left\{E\left[\widetilde{H}_k^* \widetilde{H}_k^T\right] R\right\} \;=\; tr\left\{R \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{\widetilde{H}\widetilde{H}}(e^{j2\pi f}) df\right\}.$$

By setting the gradient of J_k in (8.3) w.r.t. H_k to zero, we have

$$\begin{aligned}\left(F(q)\, Y_k Y_k^T\right)\; H_k \;&=\; F(q)\, Y_k d_k \\ &=\; F(q)\, Y_k Y_k^T H_k^o \;+\; F(q)\, Y_k n_k\,.\end{aligned}$$

Let's denote by $\widetilde{F}(q) = \dfrac{F(q)}{F(1)}$ the (dc transfer) normalized weighting window. As this window is generally low-pass, $\widetilde{F}(q)$ acts as an averaging operator, and we have

$$\widetilde{F}(q)\, Y_k Y_k^T \approx R\,.$$

On the other hand, as the optimal system variation is independent of the input signal (in the system id setup), we approximate:

$$\begin{aligned}\widetilde{F}(q) Y_k Y_k^T H_k^o \;&\approx\; \left(\widetilde{F}(q) Y_k Y_k^T\right)\; \left(\widetilde{F}(q) H_k^o\right) \\ &\approx\; R\, \widetilde{F}(q) H_k^o\,.\end{aligned}$$

Hence the filter deviation can be expressed as:

$$\widetilde{H}_k = H_k^o - H_k = \left(1 - \widetilde{F}(q)\right) H_k^o \; - \; R^{-1}\, \widetilde{F}(q) Y_k n_k$$

and the EMSE becomes:

$$\begin{aligned} EMSE \;\; &= \;\; N\sigma_n^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\widetilde{F}(e^{2j\pi f})\right|^2 df \\ &\quad + \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|1 - \widetilde{F}(e^{j2\pi f})\right|^2 tr\big\{R\, S_{HH}(e^{j2\pi f})\big\} df \end{aligned} \tag{8.5}$$

Remark that the EMSE can be broken up into two terms:

- $E_{est} = N\sigma_n^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\widetilde{F}(e^{2j\pi f})\right|^2 df$ corresponding to the estimation noise contribution; it can be interpreted as the estimation accuracy in time-invariant conditions,
- $E_{lag} \; = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|1 - \widetilde{F}(e^{2j\pi f})\right|^2 tr\left\{R\, S_{HH}(e^{j2\pi f})\right\} df$ representing the estimation error resulting from low-pass filtering the system variations (lag noise, since in the causal window case this means lagging behind).

The estimation and lag noise terms can also be interpreted as the variance and the bias of the conditional estimation problem, for a given value of the optimal filter sequence. In fact,

$$\widetilde{H}_k = \underbrace{\left(1 - \widetilde{F}(q)\right) H_k^o}_{b} \; - \; R^{-1}\, \widetilde{F}(q) Y_k n_k$$

where $b = E_{|H^o} \widetilde{H}_k$ is the estimation bias. We get

$$R_{\widetilde{H}\widetilde{H}} = E_{|H^o}\left(\widetilde{H}_k \widetilde{H}_k^T\right) = \underbrace{b\, b^T}_{\text{bias}} + \underbrace{\sigma_n^2 \left(\sum_i \tilde{f}_i^2\right) R^{-1}}_{\text{variance}} . \tag{8.6}$$

Then we see that:

- $E_{est} = N\sigma_n^2 \left(\sum_i \tilde{f}_i^2\right)$ is the variance component,
- $E_{lag} = tr\left\{R\; E_{H^o}\left[bb^T\right]\right\}$ is the bias component.

8.3 Uninformed Approach for RLS Tracking Analysis

In an uninformed approach we assume that little or no information about the system variations is available for the design of the RLS algorithm.

8.3.1 Uninformed Tracking Analysis of Causal RLS Algorithms

From (8.5), we can see that the following window characteristics characterize estimation and lag noises resp.:

- $l_\infty = \left(\sum_i \tilde{f}_i^2\right) = \left\|\widetilde{F}\right\|_2^2$ characterizing E_{est},
- $E_F(f) = \left|1 - \widetilde{F}\left(e^{j2\pi f}\right)\right|$, called *parameter tracking characteristic*, characterizing the lag noise.

To compare the tracking ability of RLS with different weighting windows, we shall choose the windows parameters such that the different algorithms behave identically under time-invariant conditions. In fact, comparing adaptive filters characterized by different values of l_∞ barely makes any sense (in the uninformed case) and it resembles "comparing runners that specialize in different distances" [144]. By normalizing the performance under time-invariant conditions, the tracking characteristic $E_F(w)$ depends only on the window shape.

Since the complexity of RLS with a rectangular window essentially doubles with respect to an exponential window, it is generally believed that this increase in complexity allows for some improvement in tracking performance. In contrast to this intuition, Fig. 8.3 shows that the plot of the normalized tracking characteristic of W-RLS lies below that corresponding to SWC-RLS. Thus, the tracking capability of W-RLS approach is better. This effect can be attributed to a higher degree of concentration of the exponential window around $i = 0$, which results in a smaller estimation delay, hence smaller bias error [144].

As we have mentioned in the Introduction, the SW-RLS approach can be generalized at a small cost for the resulting RLS algorithm to a window with three parameters (instead of just one). Compared to the Sliding and the Exponential widows, the Generalized Sliding Window introduces two extra degrees of freedom. The shape of the widow depends on the choice of these degrees of freedom. Thus, they can be optimized to minimize the average parameter tracking characteristic. In other words, the window parameters are chosen so as to

$$\begin{cases} \min\limits_{\lambda,\alpha,L} \int_0^{f_0} \left| 1 - \widetilde{F}\left(e^{j2\pi f}\right) \right|^2 df \\ \text{subject to} \sum\limits_k \tilde{f}_k^2 \;=\; l_\infty \end{cases} \tag{8.7}$$

where f_0 is the assumed bandwidth of the system variations. In Fig. 8.3, we add the tracking characteristic of the GSW-RLS estimator, as a function of $f = f_o$. As expected, the optimized GSW-RLS outperforms the SWC-RLS and W-RLS approaches.

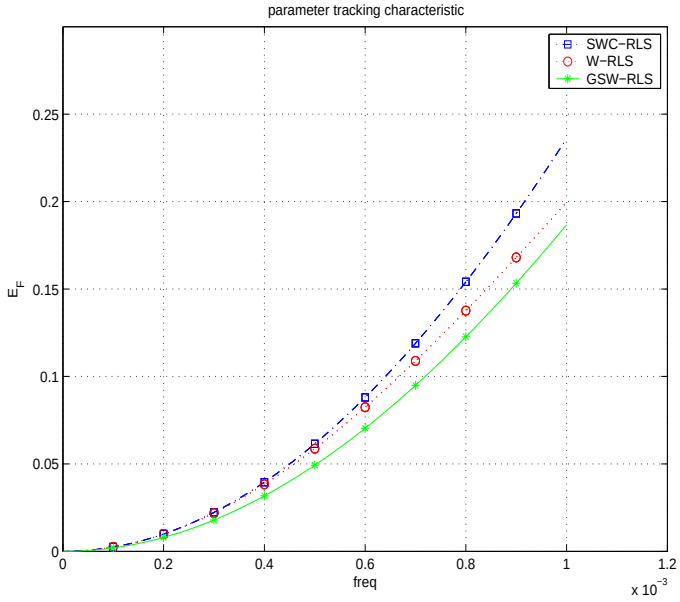

Figure 8.3: Parameter tracking characteristic for W, SWC, GSW-RLS.

8.3.2 Uninformed Tracking Analysis for Non-Causal RLS Algorithms

Bias in the RLS algorithm is caused by two kinds of distortion: amplitude and phase distortions. Phase distortion can be considerably attenuated by introducing a suitable estimation delay (using non-causal filtering). With no information about the system characteristics, a suitable estimation delay can be determined as [144]:

$$\tau_e = \sum_k k\tilde{f}_k \ , \tag{8.8}$$

the mean of the $\tilde{f}_k$ considered as a distribution. As before, under identical estimation noise, comparing the tracking capability of the non-causal RLS algorithms can be investigated by comparing what is called in [144] the *parameter matching characteristic* defined as:

$$\tilde{E}_F(w) = \left| e^{-j2\pi f\tau_e} - \widetilde{F}\left(e^{j2\pi f}\right)\right| = \left|1 - e^{j2\pi f\tau_e}\widetilde{F}\left(e^{j2\pi f}\right)\right| .$$

Fig. 8.4 shows plots of normalized matching characteristics of SWC-RLS, W-RLS, and GSW-RLS (with optimized window parameters) algorithms. The curves show that in the non-causal adaptation case, the optimal shape for the generalized window becomes the rectangular one (and in particular, rectangular windowing outperforms exponential windowing). The better parameter matching properties of the SWC-RLS approach can be explained by the linearity of the associated phase characteristic (due to the window symmetry). The delay τ_e becomes the center of the window and after delay compensation, there is zero phase distortion left.

8.4 Informed Approach for RLS Tracking Optimization

Now we suppose the statistics of the system variation to be available. In this case, we can achieve an optimal tradeoff between the estimation and lag noises. The optimal tradeoff can be found by minimizing the EMSE:

$$\min_F \left(N\sigma_n^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} df \left|F(e^{j2\pi f})\right|^2 + \int_{-\frac{1}{2}}^{\frac{1}{2}} df \left|1 - F(e^{j2\pi f})\right|^2 tr\left\{RS_{HH}(f)\right\}\right)$$

$$\text{subject to } F(1) = 1 \tag{8.9}$$

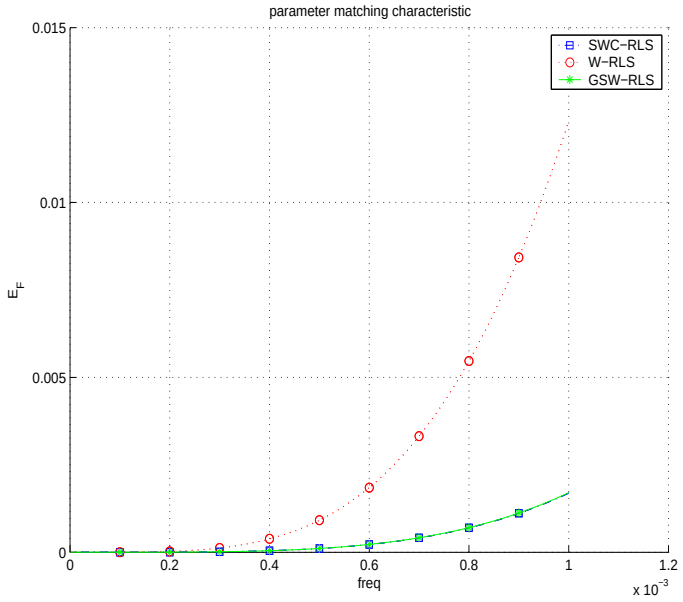

Figure 8.4: Parameter matching characteristic.

8.4.1 Optimized Causal Parametric Windows, Separable Variation Spectrum Case

In a first instance, we propose to investigate a simple (but interesting) variation model. We assume that the impulse response coefficients have proportional Doppler spectrum; i.e.

$$S_{HH}\left(e^{j2\pi f}\right) = D\, S_{hh}\left(e^{j2\pi f}\right) \ . \tag{8.10}$$

To simplify, we suppose also that the scalar spectrum S_{hh} is a flat low-pass spectrum; i.e.

$$S_{hh}\left(e^{j2\pi f}\right) = \begin{cases} 1 & |f| < f_0 \\ 0 & \text{elsewhere} \end{cases}$$

The matrix D is arbitrary but if it were diagonal (decorrelated filter coefficients) the diagonal would represent the power delay profile of the optimal filter (in wireless channel terminology). With the separable model, the Excess MSE (for the causal adaptation case) can be expressed as:

$$N\sigma_n^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} df \left|F(e^{j2\pi f})\right|^2 + tr\left\{R\, D\right\} \int_{-\frac{1}{2}}^{\frac{1}{2}} df \left|1 - F(e^{j2\pi f})\right|^2 \ .$$

The EMSE expressions for the different RLS variants become, as a function of the windows parameters:

$$
\begin{aligned}
EMSE^{SWCRLS} &= \frac{N\sigma_n^2}{L} \\
&+ 2\ tr\{R\,D\}\left(\frac{L-1}{L}f_0 - \frac{1}{\pi L^2}\sum_{k=1}^{L-1}\sin(2\pi f_0 k)\right) \\
EMSE^{WRLS} &= N\sigma_n^2\frac{1-\lambda}{1+\lambda} \\
&+ 2\ tr\{R\,D\}\left(\lambda f - \frac{\lambda}{\pi}\frac{1-\lambda}{1+\lambda}\arctan\left(\frac{1+\lambda}{1-\lambda}\tan(\pi f)\right)\right) \\
EMSE^{GSWRLS} &= N\sigma_n^2\frac{1-\lambda}{1+\lambda}\frac{1+\alpha(\alpha-2)L^{2L}}{(1-\alpha\lambda^L)^2} \\
&+ 2\ tr\{R\,D\}\left(f_0\gamma_0 + \sum_{k=1}^{\infty}\frac{\gamma_k}{\pi k}\sin(2k\pi f_0)\right)
\end{aligned}
$$

where γ_k gets computed recursively as:

$$
\begin{cases}
\gamma_{k+1} = \lambda\gamma_k - \alpha\kappa^2\lambda^{2L-1-k} & , k < L, \\
\gamma_{k+1} = \lambda\gamma_k & , k \geq L.
\end{cases}
$$

To investigate the tracking ability of the different RLS algorithms, we compare the minimum EMSE achieved by each variant(with optimized parameters). On figures (8.5) and (8.6), we gives for different values of f_0, the Excess MSE for the exponential and the rectangular widows as a function of respectively λ and L.

We notice that for the rectangular window the EMSE curve may have local minima. Therefore, we must be careful in the choice of the minimization algorithm. However, with exponential window this problem does not arise. That is why we have used the golden section algorithm to minimize the Excess MSE in the case of the exponential window ; and a multilevel quasi-exhaustive search algorithm in the case of the rectangular window. Fig. 8.7 plots the curves of minimized EMSE (as a function of the bandwidth f_0).
This analysis shows that, for a flat low-pass spectrum, the exponential window performs better than the rectangular window, but also that the optimal generalized window performs even better.

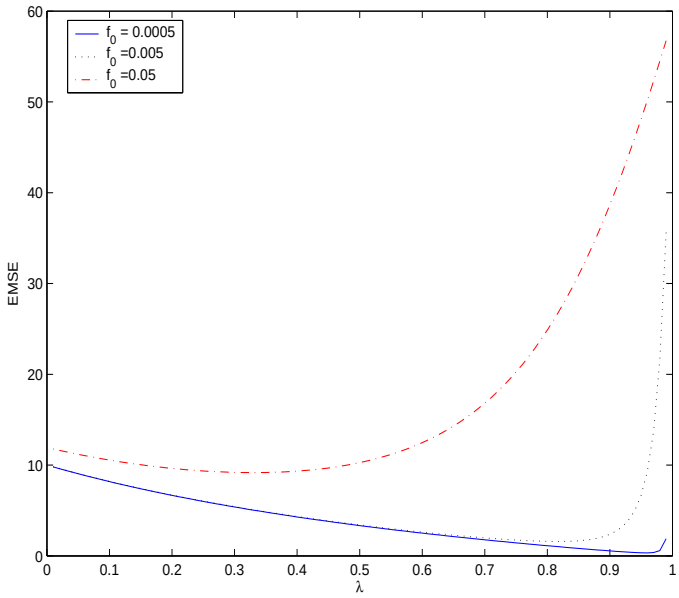

Figure 8.5: EMSE of the WRLS for diffrent values of f_0

8.4.2 Optimized Windows

Consider minimizing the Excess MSE with respect to the window coefficients. This problem can be interpreted in terms of Wiener filtering for a signal in noise problem, see Fig. 8.8, where the desired and noise signal spectra are respectively $S_{vv}\left(e^{j2\pi f}\right) = \sigma_v^2 = N\sigma_n^2$, and $S_{dd}\left(e^{j2\pi f}\right) = tr\left\{R\, S_{HH}\left(e^{j2\pi f}\right)\right\}$.

The causal Wiener solution for such problem is

$$F_{Wiener}\left(e^{j2\pi f}\right) = \frac{S_{dd}\left(e^{j2\pi f}\right)}{S_{dd}\left(e^{j2\pi f}\right) + \sigma_v^2} \,. \tag{8.11}$$

The DC component of this Wiener solution is

$$F_{Wiener}\left(1\right) = \frac{S_{dd}\left(1\right)}{S_{dd}\left(1\right) + \sigma_v^2} \,. \tag{8.12}$$

Now, since S_{dd} is quite lowpass , we have $S_{dd}(1) >> \sigma_v^2$. Thus, for an acceptable SNR, $F_{Wiener}(1) \approx 1$. Then,

$$F_{opt}\left(e^{j2\pi f}\right) \approx F_{Wiener}\left(e^{j2\pi f}\right) = \frac{S_{dd}\left(e^{j2\pi f}\right)}{S_{dd}\left(e^{j2\pi f}\right) + \sigma_v^2} \,. \tag{8.13}$$

The associated Excess MSE is given by:

$$EMSE_{min} = \sigma_v^2 \int F_{opt}\left(e^{j2\pi f}\right) df = \sigma_v^2\, f_0^{opt} \,. \tag{8.14}$$

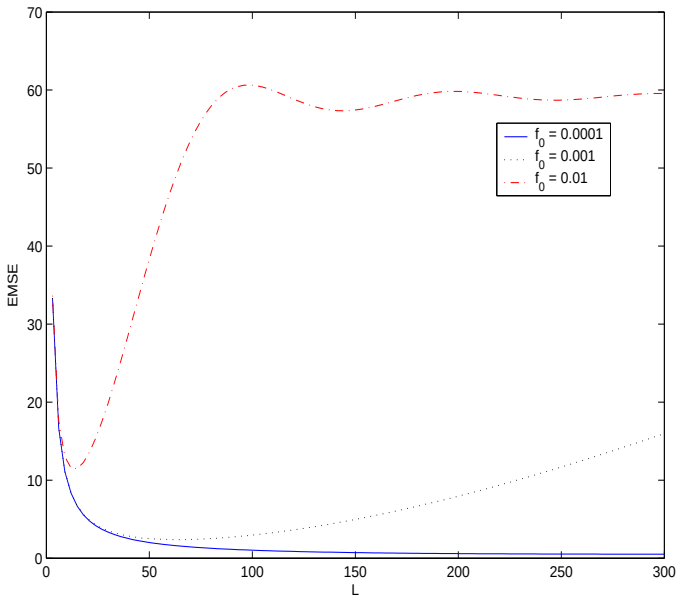

Figure 8.6: EMSE of the SWC RLS for diffrent values of f_0

If we impose a causality constraint, the Wiener solution becomes:

$$F^c_{Wiener}\left(e^{j2\pi f}\right) = 1 - \frac{\sigma_v^2}{\sigma^2 A\left(e^{j2\pi f}\right)} \tag{8.15}$$

where $A(z)$ denotes the optimal prediction error filter for the signal $x = d + v$, and σ^2 the associate prediction error variance. Using the same arguments as before, one can show that, for an acceptable SNR, $F^c_{Wiener}(1) \approx 1$; and then $F^c_{opt}\left(e^{j2\pi f}\right) \approx F^c_{Wiener}\left(e^{j2\pi f}\right)$. The associated Excess MSE is given by:

$$EMSE^c_{min} = \sigma_v^2 \int F^c_{opt}\left(e^{j2\pi f}\right) df = \sigma_v^2 \, f_0^{c,opt}. \tag{8.16}$$

8.4.3 Optimality Considerations for Classical Windows

The question we investigate in this section is "For which optimal filter variation model is the exponential or the rectangular window optimal?". To answer this question, we use the reverse engineering technique. We assume a separable optimal filter variation spectrum with uncorrelated optimal filter coefficients.

Equating the causal Wiener solution and the exponential window transfer function

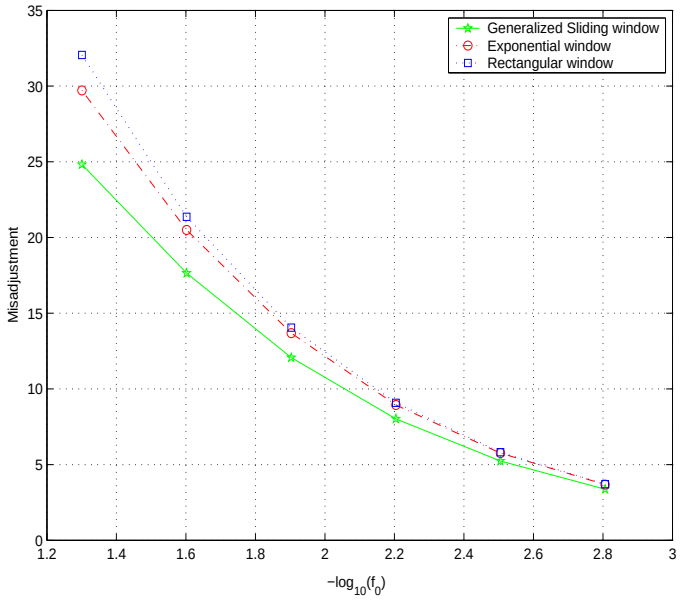

Figure 8.7: $EMSE_{min}$ curves for flat low-pass variations.

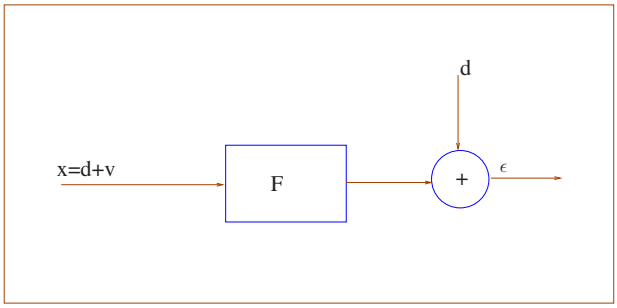

Figure 8.8: Signal in noise problem

leads to a variation spectrum of the form:

$$S_{hh}(z) \propto \frac{1}{(1-\alpha z^{-1})\,(1-\alpha z)}\ . \tag{8.17}$$

Thus, for AR(1) "drifting" parameters, the exponential window optimizes the tracking performance (over the set of all causal windows).

Similarly, by equating the non-causal Wiener filter with the centered rectangular window, one can show that SWC-RLS is optimal for a variation spectrum of the form:

$$S_{hh}(z) \propto \frac{F_R(z)}{1-F_R(z)} \tag{8.18}$$

where $F_R(z) = \frac{1}{L}\frac{z^{\frac{L}{2}} - z^{-\frac{L}{2}}}{z^{\frac{1}{2}} - z^{-\frac{1}{2}}}$ denotes the centered sliding window transfer function. $S_{hh}(z)$ in (8.18) can be interpreted to be the spectrum of a process g with spectrum $S_{gg}(z) = F_R(z)$), low-pass filtered by $H(z) = \left(\frac{1}{1-F_R(z)}\right)^{1/2}$ (spectral factor). The input g can be interpreted as a white jumping process (with memory $L/2$), or as white noise viewed on the scale L/2: g stays constant for $L/2$ samples, then jumps to an uncorrelated value, stays constant again for $L/2$ samples and so on. On the other hand, around frequency zero

$$\frac{1}{1 - F_R\left(e^{2j\pi f}\right)} \approx \frac{3}{\pi^2(L^2-1)}\,\frac{1}{f^2}\,. \tag{8.19}$$

Then, the low-pass filter $H(z)$ can be interpreted as an integrator.

8.5 Concluding Remarks

In this chapter, we have considered tracking of a time-varying system modeled as a stationary vector process. In the Uninformed approach, we have investigated the tracking capability, by comparing the tracking and matching characteristics of the different RLS windows (under identical estimation noise). In the Informed approach, we have interpreted the RLS algorithm as a filtering operation on the optimal filter process and the instantaneous gradient noise. We have shown the optimality of the exponential window for AR(1) "drifting" parameters, and of the sliding window for integrated white "jumping" parameters. An open question remains: how to estimate and optimize simultaneously the adaptive filter and window parameters? Alternatively, one may opt for a two-step approach:

- Step 1: non-causal SWC-RLS (with short window), providing noisy but undistorted filter estimates.
- Step 2: Wiener (Kalman) filtering to provide the optimal estimation noise/low-pass distortion compromise, e.g. as in [145].

Such an approach allows for less constrained optimal filtering, that can be optimized, and tailored to individual (and possibly correlated) filter coefficients, whereas RLS has only one global window.

APPENDIX

8.6 EMSE for a Causal Rectangular Window, Separable Variation Spectrum Case

The z-transform of the rectangular window is given by:

$$F_R(z) = \frac{1}{L}\,\frac{1-z^{-L}}{1-z^{-1}} \tag{8.20}$$

In order to perform the analytical expression of the Excess MSE in the case of a rectangular window, we start calculating the quantity $\int_0^{f_0} \left|1 - F_R(e^{2j\pi f})\right|^2 df$.

$$\begin{aligned}\left|1 - F_R(e^{2j\pi f})\right|^2 &= \left|\frac{L - Le^{-2j\pi f} - 1 + e^{-2j\pi fL}}{L\left(1 - e^{-2j\pi f}\right)}\right|^2 \\ &= \frac{\left|(L-1) - Le^{-2j\pi f} + e^{-2j\pi fL}\right|^2}{4L^2 \sin^2(\pi f)}\end{aligned}$$

The expansion of the numerator gives:

$$\begin{aligned}\left|(L-1) - Le^{-2j\pi f} + e^{-2j\pi fL}\right|^2 &= 4L(L-1)\sin^2(\pi f) \\ &+4L\sin^2(\pi f(L-1)) - 4(L-1)\sin^2(\pi fL)\end{aligned}$$

Finally, we have:

$$\begin{aligned}\left|1 - F_R(e^{2j\pi f})\right|^2 &= \frac{L-1}{L} + \frac{1}{L}\frac{\sin^2(\pi f(L-1))}{\sin^2(\pi f)} \\ &- \frac{L-1}{L^2}\frac{\sin^2(\pi fL)}{\sin^2(\pi f)}\end{aligned}$$

We can see, from the previous expression, that $\int_0^{f_0} \left|1 - F_R(e^{2j\pi f})\right|^2 df$ can be deduced from the function

$$g(f_0) = \int_0^{f_0} \frac{\sin^2(\pi fL)}{\sin^2(\pi f)} df$$

The Fejer kernel can be written as

$$\frac{\sin(\pi fL)}{\sin(\pi f)} = e^{j\pi f(L-1)}\left(\sum_{k=0}^{L-1} e^{-2j\pi fk}\right)$$

Then,

$$\frac{\sin^2(\pi f L)}{\sin^2(\pi f)} = L + 2\sum_{k=1}^{L-1}(L-k)\cos(2\pi f k)$$

Finally,

$$\int_{-f_0}^{f_0} \left|1 - F_R(e^{2j\pi f})\right|^2 df = 2\frac{L-1}{L}f_0 - \frac{2}{\pi L^2}\sum_{k=1}^{L-1}\sin(2\pi f_0 k)$$

$*$ CONCLUSION

Using a rectangular windowing, The Excess MSE is given by:

$$\begin{aligned} EMSE &= \frac{N\sigma_n^2}{L} \\ &+2tr(RD)\left(\frac{L-1}{L}f_0 - \frac{1}{\pi L^2}\sum_{k=1}^{L-1}\sin(2\pi f_0 k)\right) \end{aligned} \tag{8.21}$$

8.7 EMSE for a Causal Exponential Window, Separable Variation Spectrum Case

The z-transform of the exponential window is given by:

$$F_E(z) = \frac{1-\lambda}{1-\lambda z^{-1}} \tag{8.22}$$

In order to perform the analytical expression of the Excess MSE in the case of a exponential window, we start calculating the quantity $\int_0^{f_0} \left|1 - F_E(e^{2j\pi f})\right|^2 df$.

$$\left|1 - F_E(e^{2j\pi f})\right|^2 = \frac{2\lambda^2(1-\cos(2\pi f))}{1-2\lambda\cos(2\pi f)+\lambda^2}$$

The previous expression can be expanded as

$$\begin{aligned} \left|1 - F_E(e^{2j\pi f})\right|^2 &= \frac{2\lambda^2 - 2\lambda}{1-2\lambda\cos(2\pi f)+\lambda^2} \\ &+ \frac{2\lambda(1-\lambda\cos(2\pi f))}{1-2\lambda\cos(2\pi f)+\lambda^2} \end{aligned} \tag{8.23}$$

On the other hand, we have:

$$\begin{aligned}\int \frac{1-a^2}{1-2a\cos(x)+a^2}dx &= 2\text{arctan}\left(\frac{1+a}{1-a}\tan\left(\frac{x}{2}\right)\right)\\ \int \frac{(1-a\cos x)}{1-2a\cos x+a^2}dx &= \frac{x}{2}+\text{arctan}\left(\frac{1+a}{1-a}\tan\left(\frac{x}{2}\right)\right)\end{aligned}$$

Using the integration change of variables $x = 2\pi f$, we have

$$\begin{aligned}\int_0^{f_0}\left|1-F_E(e^{2j\pi f})\right|^2 df &= \lambda f_0\\ &\quad -\frac{\lambda}{\pi}\frac{1-\lambda}{1+\lambda}\text{arctan}\left(\frac{1+\lambda}{1-\lambda}\tan\left(\pi f_0\right)\right)\end{aligned}$$

∗ CONCLUSION

Using an exponential windowing, The Excess MSE is given by:

$$\begin{aligned}EMSE &= N\sigma_n^2\frac{1-\lambda}{1+\lambda}\\ &\quad +2tr(RD)\left(\lambda f_0-\frac{\lambda}{\pi}\frac{1-\lambda}{1+\lambda}\text{arctan}\left(\frac{1+\lambda}{1-\lambda}\tan\left(\pi f_0\right)\right)\right)\end{aligned} \tag{8.24}$$

8.8 EMSE for a Causal Generalized Sliding Window, Separable Variation Spectrum Case

The z-transform of the generalized sliding window is given by:

$$F_G(z)=\kappa\frac{1-\alpha\lambda^L z^{-L}}{1-\lambda z^{-1}} = \frac{1-\lambda}{1-\alpha\lambda}\frac{1-\alpha\lambda^L z^{-L}}{1-\lambda z^{-1}} \tag{8.25}$$

In order to perform the analytical expression of the Excess MSE in the case of a generalized sliding window, we start calculating the quantity $\int_0^{f_0}\left|1-F_G(e^{2j\pi f})\right|^2 df$.

By expanding $\left|1-F_G(e^{2j\pi f})\right|^2$, we have:

$$\begin{aligned}\left|1-F_G(e^{2j\pi f})\right|^2 &= A(\alpha,\lambda,L)\frac{1}{1+2\lambda\cos(2\pi f)+\lambda^2}\\ &+ B(\alpha,\lambda,L)\frac{\cos(2\pi f)}{1+2\lambda\cos(2\pi f)+\lambda^2}\\ &+ C(\alpha,\lambda,L)\frac{\cos(2\pi fL)}{1+2\lambda\cos(2\pi f)+\lambda^2}\\ &+ D(\alpha,\lambda,L)\frac{\cos(2\pi f(L+1))}{1+2\lambda\cos(2\pi f)+\lambda^2}\end{aligned}$$

where

$$-A(\alpha,\lambda,L) = \frac{\left(\lambda-\alpha\lambda^L\right)^2+\lambda^2\left(1-\alpha\lambda^L\right)^2}{\left(1-\alpha\lambda^L\right)^2} + \frac{\alpha^2\lambda^{2L}(1-\lambda)^2}{\left(1-\alpha\lambda^L\right)^2}$$

- $B(\alpha,\lambda,L) = \dfrac{2\lambda\left(\lambda-\alpha\lambda^L\right)}{\left(1-\alpha\lambda^L\right)}$
- $C(\alpha,\lambda,L) = \dfrac{\lambda^L(1-\lambda)\left(\lambda-\alpha\lambda^L\right)}{\left(1-\alpha\lambda^L\right)^2}$
- $D(\alpha,\lambda,L) = \dfrac{\alpha\lambda^{L+1}(1-\lambda)}{\left(1-\alpha\lambda^L\right)}$

Thus, to evaluate $\int_0^{f_0}\left|1-F_G(e^{2j\pi f})\right|^2 df$, it is sufficient to have the expression of

$$\int_0^{f_0}\frac{\cos(2\pi fk)}{1+2\lambda\cos(2\pi f)+\lambda^2}df \qquad k\geq 0,\ \ 0<\lambda<1$$

Other approach to evaluate, approximately, the previous expression is to expend it on a Fourier series.

Let us define the sequence $\{\widetilde{w}_k\}_k$ as:

$$\begin{cases}\widetilde{w}_0 = 1-w_0\\ \widetilde{w}_k = -w_k\end{cases} \tag{8.26}$$

Using the previous notations, we write

$$1-F_G\left(e^{2j\pi f}\right) = \sum_{k=0}^{\infty}\widetilde{w}_k e^{-2jk\pi f} \tag{8.27}$$

Then,

$$\left|1-F_G\left(e^{2j\pi f}\right)\right|^2 = \gamma_0 + 2\sum_{k=0}^{\infty}\gamma_k\cos(2k\pi f)$$

where $\gamma_k = \sum_{p=0}^{\infty}\widetilde{w}_p\widetilde{w}_{p+k}$. Note that γ can be calculated recursively using:

$$\begin{cases}\gamma_{k+1} = \lambda\gamma_k - \alpha\kappa^2\lambda^{2L-1-k} & k<L\\ \gamma_{k+1} = \lambda\gamma_k & k\geq L\end{cases} \tag{8.28}$$

Let us define the function series

$$f_k(x) \ : \ x \longmapsto \gamma_k \cos\left(2k\pi f\right)$$

we verify that:

$$|f_k(x)| \leq \text{Cst} \ \ \lambda^k$$

Thus, we show that $\sum_{k=0}^{\infty} f_k$ is normally, then uniformly, convergent. What's imply that $\int_0^{f_0} f_k dx$ is a summable sequence, and we have

$$\int_0^{f_0} \sum_{k=0}^{\infty} f_k \ = \ \sum_{k=0}^{\infty} \int_0^{f_0} f_k$$

Thus,

$$\int_0^{f_0} \left|1 - F_G\left(e^{2j\pi f}\right)\right|^2 df = f_0\gamma_0 \ + \ \sum_{k=1}^{\infty} \frac{\gamma_k}{\pi k} \sin\left(2k\pi f_0\right) \tag{8.29}$$

$*$ CONCLUSION

Using a generalized slinding windowing, The Excess MSE is given by:

$$\begin{aligned} EMSE \ &= \ N\sigma_n^2 \frac{1-\lambda}{1+\lambda} \frac{1+\alpha(\alpha-2)\lambda^{2L}}{\left(1-\alpha\lambda^L\right)^2} \\ &+ \ 2tr(RD) \left(f_0\gamma_0 \ + \ \sum_{k=1}^{\infty} \frac{\gamma_k}{\pi k} \sin\left(2k\pi f_0\right)\right) \end{aligned} \tag{8.30}$$

We have shown that the EMSE can be written as an infinite sum of a summable sequence. Then, we can derive an EMSE approximation by truncating the summation in a given order. As γ_k decreases exponentially, we can verify that the estimation error decreases, also, exponentially as:

$$\left|\widehat{EMSE}_K - EMSE\right| \leq Cst \ \lambda^K \tag{8.31}$$

where $\widehat{EMSE}_K$ denotes the K-order approximation of the EMSE.

Chapter 9

Conclusions and Perspectives

In the first part of this thesis we have tackled the problem of Bayesian Adaptive Filtering on non-stationary environment. Thus in chapter 5, we first outlined the existing adaptive filtering technics and highlighted their limitations. We then provided a first BAF technique based on the LMMSE solution (Wiener Filtering). We tested our proposed approach on mobile radio channel and the comparison results with standard adaptive filtrers showed the good offered performances.

The work in chapter 6 was motivated by the well to provide applicable algorithms with performance approaching those of the optimal case. In fact, the only existing optimal approach is the Kalman filter, in which the time-varying optimal filter is modeled as a vector AR(1) process. The Kalman filter is in practice never applied as an adaptive filter because of its complexity and large number of unknown parameters in its state-space (AR(1)) model. In this chapter we look for practical techniques to take advantage of the Kalman optimality with reduced complexity to make this approach applicable. We thus propose and algorithm based on modeling the optimal adaptive filter coefficients as a stationary vector process, in particular as a AR(1) model which results on a Kalman filter where the model parameters are adapted using EM. This algorithm has complexity $O(N^3)$ which we reduce to $O(N^2)$ (complexity of RLS) by proposing a diagonal AR(1) based model. To further reduce the complexity, we provide a second technique called component-wise EM-Kalman (complexity $O(N)$ comparable to LMS). We also give the exact analytical expressions of EMSE in the steady-state in the general case of all proposed algorithms.

The techniques are compared for a radio mobile communication scenario.
In chapter 7, we develop an other technique in order to obtain the performance as the EM-Kalman with complexity $O(N)$.
The second part of the thesis focuses on the study of windoinw optimization of RLS algorithm. Thus in chapter 8, we consider the tracking of a time-varying system modeled as a stationary vector process. In the case where no information on the channel statistics is available, we investigated the tracking capability, by comparing the tracking and matching characteristics of the different RLS windows (under identical estimation noise). In the case where such an information is available, we interpreted the RLS algorithm as a filtering operation on the optimal filter process and the instantaneous gradient noise. We was interested on finding the model for which the exponential and the rectangular windows are optimal. We thus showed the optimality of the exponential window for AR(1) "drifting" parameters, and of the sliding window for integrated white "jumping" parameters.

Areas for further research

If complexity is not an issue but only performance counts: KF: need to develop AR(∞) state equation and colored measurement noise and average resulting $P_{k|k}$ w.r.t. input signal.

How about staying close to the Kalman Filter, given input signal. Does this make sense for the non-sytem identification applications? But nonetheless, suppose don't care about complexity and just want the best solution.

Issue of smoothing versus filtering and prediction, influence of estimation delay on perf. Solution for arbitrary (positive: prediction, or negative: smoothing) delay based on the prediction error filter for the adaptive filter estimates. Prediction approaches don't require additive noise variance.

Issue of state transition matrix: diagonal of phases in complex case or block diagonal of 2x2 rotations in real case (frequency shifted/non-centered Doppler spectra)?

Kalman filter type solution for BEM subsampled model? Multi-rate KF.

how to estimate and optimize simultaneously the adaptive filter and window parameters? Alternatively, one may opt for a two-step approach:

- Step 1: non-causal SWC-RLS (with short window), providing noisy but undistorted filter estimates.
- Step 2: Wiener (Kalman) filtering to provide the optimal estimation noise/low-pass distortion compromise, e.g. as in [145].

Such an approach allows for less constrained optimal filtering, that can be optimized, and tailored to individual (and possibly correlated) filter coefficients, whereas RLS has only one global window.

Bibliography

[1] S. Haykin, ”Adaptive Filter Theory,” Prentice Hall, 4th edition, 2001

[2] J. G. Proakis, ”Digital Communications,” Mc-Graw Hill, 2001.

[3] V. Solo and X. Kong, ”Adaptive Signal Processing Algorithms: Stability and Performance,” Prentice Hall, 1994

[4] A. Benveniste and M. Métivier and P. Priouret, ”Adaptive Algorithms and Stochastic Approximations,” Springer-Verlag, Berlin, 1990

[5] L. Ljung and T. Soderstrom, ”Theory and Practice of Recursive Identification,” MIT Press, Cambridge, MA, 1983

[6] M. Niedzwiecki, ”Identification of Time-Varying Systems,” Wiley, 2000

[7] B. D. O. Anderson and J.B. Moore, ”Optimal Filtering,” Prentice-Hall, Englewood Cliffs, NJ, 1979

[8] T. S. Rappaport, ”Wireless Communications: Principles and Practice,” Second edition, Prentice Hall, 2002.

[9] M. Najim, ”Digital Filters Design for Signal And Image Processing,” *Iste Publishing Company,*, novembre, 2005.

[10] D. T. M. Slock, “On the Convergence Behavior of the LMS and the Normalized LMS Algorithms,” *IEEE Trans. on Signal Processing,* Vol. 41, pp. 2811-2825, No. 9, Sept. 1993.

[11] S. S. Kozat and A. C. Singer, "Further Results in Multistage Adaptive Filtering" *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Proc.,* March 2002.

[12] T. Sadiki and D. T. M. Slock, "Bayesian Adaptive Filtering At linear Cost" *IEEE Workshop Statistical signal Processing,* Bourdaux, France, Jul. 2005.

[13] S. Haykin and A. H. Sayed and J. R. Zeidler and P. Yee and P. C. Wei, "Adaptive tracking of linear time-variant systems by extended RLS algorithms," *IEEE Trans. on Signal Processing,* Vol. 45, pp. 1118–1128, No. 5, May 1997.

[14] Lei Guo and Lennart Ljung, "Performance Analysis of General Tracking Algorithms," *IEEE Trans. on Automatic Control,* Vol. 40, No. 8, 1995.

[15] L. Vandendorpe, "Multi-tone Spread Spectrum Communications System in an Indoor Wireless Channel," *Proc. of First IEEE Symposium on Communications and Vehicular Technology in the Benelux,* pp. 4.1-1/4.1-8, Oct. 1993.

[16] A. Benveniste, "Design of Algorithms for Tracking of Time-Varying Systems," *Int.J Adaptive Contr. Sig. Proces.,* Vol. 1, pp. 3–29, 1987.

[17] L. Guo, "Estimation Time-Varying Parameters by Kalman Filter Based Algoritmis: Stability and Convergence," *Int.J Adaptive Contr. Sig. Proces.,* Vol. 35, pp. 141–147, 1990.

[18] L. Guo, L. Ljung, "Performance Analysis of the Forgetting Factor RLS Algorithm," *Int. J. Adaptive Cont. Sig. Proces.,* Vol. 7, pp. 141–537, 1993.

[19] L. Ljung and S. Gunnarsson, "Adaptive Tracking in System Identification, A Survey," *Automatica.,* Vol. 26, No. 1, pp. 7–22, 1990.

[20] J. B. Evans and P. Xue and B. Liu, "Analysis and Implementation of Variable Step Size Adaptive Algorithms," *IEEE Trans. Signal Proc.,* Vol. 41, No. 8, pp. 2517–2535, 1993.

[21] E. Eleftheriou and D. Falconer, "Tracking Properties and Steady-State Performance of RLS Adaptive Filter Algorithms," *IEEE Trans. ASSP.,* Vol. 34, No. 5, pp. 1097-1110, Oct 1986.

[22] B. Widrow et al., "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter," *IEEE Trans. Signal Proc.,* Vol. 64, No. 8, pp. 1151-1162, Aug. 1976.

[23] O. M. Macchi and N. J. Bershad, "Adaptive Recovery of a Chirped Sinusoid in Noise, Part I: Performance of the RLS Algorithm, Part II: Performance of the LMS Algorithm" *IEEE Trans. Signal Proc.,* Vol. 39, , pp. 583-602, March 1991.

[24] M. Niedzwiecki and T. Klaput, "Fast Recursive Basis Function Estimators for Identification of Time-Varying Processes," *IEEE Trans. Signal Proc.,* Vol. 50, No. 8, Aug. 2002.

[25] M. Wittmann, J. Marti, T. Kurner, "Impact of the Power Delay Profile Shape on the Bit Error Rate in Mobile Radio Systems," *IEEE Trans. on Vehicular Technology,* Vol. 46, No. 2, May 1997.

[26] J. Hansen, "An Analytical Calculation of Power Delay Profile and Delay Spread With Experimental Verification," *IEEE Communications Letters,* Vol. 7, No. 6, June 2003.

[27] J. Lavergnat and P. Gole, "Statistical behavior of a simulated microwave multipath channel," *IEEE Trans. Antennas Propagat.,* vol. 39, No. 12, pp. 1697-1706, 1991.

[28] I. Crohn and E. Bonek, "Modeling of intersymbol interference in a Rayleigh fast fading channel with typical power delay profiles," *IEEE Trans. on Vehicular Technology,* Vol. 35, No. 4, pp. 438-447, 1992.

[29] F. Davarian, "Channel simulation to facilitate mobile-satellite communications research," *IEEE Trans. Commun.,* Vol. 35, No. 1, pp. 47-56, 1987.

[30] D. Parsons, "The Mobile Radio Propagation Channel," London: Pentech, 1992.

[31] E. Lutz, D. Cygan, M. Dippold, F. Dolansky, and W. Papke, "The land mobile satellite communication channel-recording, statistics and channel model," *IEEE Trans. Veh. Technol.,* vol. 40, no. 2, pp. 375-385, 1991.

[32] M. Patzold, A. Szczepanski, N. Youssef, "Methods for modeling of specified and measured multipath power-delay profiles," *IEEE Tranis. Veh. Tech.,* vol. 51, pp. 978-988, Sep. 2002.

[33] J. C. I Chuang, "The effects of time delay spread on portable radio communications channels with digital modulation," *IEEE J. Select. Areas Common.,* vol 5, pp. 879-889, Sep. 1987.

[34] P. A. Bello, "Characterization of Randomly Time-Variant Linear Channels," *IEEE Transactions on Communications Systems,* vol. 11, pp. 360-393, Dec. 1963.

[35] W. R. Braun and U. Dersch, " A physical mobile radio channel model," *IEEE Trans. Veh. Technol.,* vol. 40, no. 2, pp. 472 482, 1991.

[36] D. L. Nielson, "Microwave propagation measurements for mobile digital radio application," *IEEE Trans. Veh. Technol.,* vol. VT-27, no. 3, pp. 117 132, 1978.

[37] R. K. Raymond and W. J. Edward, "A Variable Step Size LMS Algorithm,"' *IEEE Trans. Signal Proc.,* Vol. 40, No. 7, Jul. 1992.

[38] V. J. Mathews and Z. Xie, "A Stochastic Gradient Adaptive Filter with Gradient Adaptive Step Size,"' *IEEE Trans. Signal Proc.,* Vol. 41, No. 6, pp.2075–2087, June. 1993.

[39] J. B. Evans and P. Xue and B. Liu, "Analysis and Implementation of Variable Step Size Adaptive Algorithms,"' *IEEE Trans. Signal Proc.,* Vol. 41, No. 8, pp.2517–2535, Aug. 1993.

[40] W. Liu, "Performance of Joint Data and Channel Estimation Using Tap Variable Step-Size (TVSS) LMS for Multipath Fast Fading Channel," *IProc. Globecom,* pp.973–978, 1994.

[41] G. L. Cassio and M. B. Jose Carlos, "Evaluation and Design of Variable Step Size Adaptive Algorithms," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Proc,* Vol. May 1994.

[42] T. Aboulnasr and K. Mayyas, "A Robust Variable Step-Size LMS-Type Algorithm: Analysis and Simulations," *IEEE Trans. on Signal Proc.,* Vol. 45, No. 3,March 1997.

[43] T. Kailath, "Linear Systems," Prentice Hall, New Jersey, 1980.

[44] T. McKelvey, " Discussion, on the use of minimal parametrizations in multivariable ARMAX identification" by R.P. Guidorzi, European Journal of Control, 4 (1998), pp. 93-98.

[45] T. McKelvey, "System identification using an over-parametrized model class: Improving the optimization algorithm," *in Proceedings of the 36th IEEE Conference on Decision and Control,* San Diego, Ca, 1997, pp. 2984-2989.

[46] L. Ljung,"System Identification: Theory for the User", *Prentice-Hall, Inc.,* New Jersey, 2nd ed., 1999.

[47] L. Ljung"MATLAB System Identification Toolbox Users Guide," Version 5, The Mathworks, 2000.

[48] T. Soderstrom and P. Stoica, "System Identification," *Prentice Hall,* New York, 1989.

[49] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE Transactions on Automatic Control,* Vol. 23 , pp. 770-783, 1978 .

[50] L. Ljung and P. Caines, "Asymptotic normality of prediction error estimators for approximate system models," Stochastics, 3 (1979), pp. 29 46.

[51] R. Fisher, "On an absolute criterion for fitting frequency curves," Mess. Math., 41 (1912), p. 155.

[52] G. Box and G. Jenkins, "Time Series Analysis, Forecasting and Control," Holden-Day, San Francisco, 1970.

[53] E. Hannan, "Multiple Time Series", Wiley, New York, 1970.

[54] A. H. Jazwinski, "Stochastic Processes and Filtering Theory," Academic Press, 1970.

[55] G. Goodwin and R. Payne, "Dynamic System Identification," Academic Press, 1977.

[56] S. Zacks, "The Theory of Statistical Inference," Wiley, New York, 1971.

[57] H. Cramer, "Mathematical Methods of Statistics," Princeton Univ. Press, Princeton, New Jersey, 1946.

[58] R. Kohn, "Asymptotic properties of time-domain Gaussian estimators," Adv. Appl. Prob, 10 (1978), pp. 339-359.

[59] J. Dennis and R. B. Schnabel, "Numerical Methods for Unconstrained Optimization and Nonlinear Equations," Prentice Hall, 1983.

[60] A. van Overbeek and L. Ljung, "On-line structure selection for multivariable state-space models," Automatica, 18 (1982), pp. 529-543.

[61] P. Kabaila and G. Goodwin, "On the estimation of the parameters of an optimal interpolator when the class of interpolators is restricted," *SIAM Journal of Control and Optimisation,* 18 (1980), pp. 615-622.

[62] R. Jennrich, "Asymptotic properties of nonlinear least squares estimators," *Annals of Mathematical Statistics,* 40 (1969), pp. 633-643.

[63] W. Larimore, "Canonical variate analysis in identification, filtering and adaptive control," *in Proceedings of the 29th IEEE Conference on Decision and Control,* Hawaii, 1990, pp. 596-604.

[64] S. B. Gelfand and Y. Wei and V. K. James, "The Stability of Variable Step-Size LMS Algorithms," *IEEE Transactions on Signal Proc. Theory,* Vol. 47, Dec. 1999.

[65] C. Rusu and C. F. N. Cowan, "Convex Variable Step-Size (CVSS) Algorithm," *IEEE Transactions on Signal Proc. Letters Theory,* Vol. 7, No. 9 Sep. 2000.

[66] R. C. Bilcu and P. Kuosmanen and K. Egiazarian, "A Transform Domain LMS Adaptive Filter with Variable Step-Size," *IEEE Transactions on Signal Proc. Letters Theory,* Vol. 9, No. 2 Feb. 2002.

[67] P. Billingsley, "Probability and Measure," Wiley , New York, 1986.

[68] S. S. Kozat and A. C. Singer,"Further Results in Multistage Adaptive Filtering," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Proc.,* March. 2002.

[69] J. Arenas-Garcia and V. Gomez-Verdejo and M. Martinez-Ramon and A. R. Figueiras-Vidal "Separate-Variable Adaptive Combination of LMS Adaptive Filters for Plant Identification," *Proc. IEEE NNSP Workshop,* Toulouse, France, Sept. 2003.

[70] Chaer, W. S. and Bishop, R. H. and Ghosh, J. "A Mixture-of-Experts Framework for Adaptive Kalman Filtering," *IEEE Trans. Systems, Man and Cybernetics, Part B,* Vol. 27, No. 3 , June 1997.

[71] W. J. Song and M. S. Park, "A Comlementary Pair LMS Algorithm for Adaptive Filtering," *Proc. ICASSP,* Münich, Germany, Apr. 1997.

[72] L. Lindbom and J. Rutstron and A. Ahlen and M. Sternad, "Automatic Tuning of Step Size in WLMS Algorithms: Application to {EDGE}," *Proc. IEEE VTC fall,* 2002.

[73] M. Sternad and L. Lindbom and A. Ahlen, "Robust Wiener Design of Adaptation Laws Constant Gains," *Proc. IFAC Workshop on Adaptation and Learning in Control and Signal,* Como, Italy, Aug. 2001.

[74] L. Lindbom and M. Sternad and A. Ahlen, "Adaptation with Constant Gains: Analysis for Fast Variations," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Proc.,* Orlando,FL, 2002.

[75] L. Lindbom and A. Ahlèn and M. Sternad and M. Falkenstrom, "Tracking of Time-Varying Mobile Radio Channels Part II: A Case Study," *IEEE Trans. on Communication,* Jan. 2002.

[76] L. Lindbom and A. Ahlèn and M. Sternad and M. Falkenstrom, "Tracking of Time-Varying Mobile Radio Channels Part I: The Wiener LMS Algorithm," *IEEE Trans. on Communication,* Vol. 49 No. 12, Dec. 2001.

[77] L. Lindbom and A. Ahlèn and M. Sternad and M. Falkenstrom, "Wiener Design of Adaptation Algorithm with Time-Invariant Gains," *IEEE Trans. on Signal Proc.,* Vol. 50 No. 8, Aug. 2002.

[78] R. K. Martin and C. R. Johnson, "NSLMS: a Proportional Weight Algorithm for Sparse Adaptive Filters," *Proc. Asilomar Conf. Signals Systems and Computers,* Pacific Grove, CA, USA, 2001.

[79] J. Benesty and S. L. Gay, "An Improved PNLMS Algorithm," *Proc. ICASSP,* Orlando, FL,USA, 2002.

[80] J. Benesty and T. Gänsler and D. R. Morgan and M. M. Sondhi and S. L. Gay, "Advances in Network and Acoustic Echo Cancellation," Springer, chapter, Adaptive Proportionate Step-Size Algorithms, 2001.

[81] D. T. M. Slock, "Fractionally-Spaced Subband and Multiresolution Adaptive Filters," *Proc. ICASSP,* Toronto, Canada, May 1991.

[82] D. T. M. Slock, "Fast Transversal Filters with Data Sequence Weighting," *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol. 37, No. 3, pp. 346–359, March 1989.

[83] W. P. Ang and H. K. Garg, "A New Adaptive Channel Etimator For Turbo Decoding in Rayleigh Fading Channel," *Proc. 8th Int'l Conf. Communication Systems (ICCS),* 2002.

[84] D. Schafhuber and G. Matz and F. Hlawatsch and P. Loubaton, "MMSE Estimation of Time-Varying Channels for DVB-T Systems with Strong Co-Channel Interference," *Proc. European Signal Processing Conference (EUSIPCO),* Toulouse, France, Sept. 2002.

[85] M. Lenardi and D. T. M. Slock, "Estimation of Time-Varying Wireless Channels and Application to the UMTS W-CDMA FDD Downlink," *Proc. European Wireless (EW),* Florence, Italy, Feb. 2002.

[86] G. Montalbano and D. T. M. Slock, "Joint Common-Dedicated Pilots Based Estimation of Time-Varying Channels for W-CDMA Receivers," *Proc. Vehic. Tech. Conf. (VTCfall),* Orlando, FL, USA, Sept.. 2003.

[87] H. Akaike, "On the use of a linear model for the identification of feedback systems," *Ann. Inst. Stat. Math.,* 20 (1968), pp. 425-439.

[88] H. Akaike, "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika,* 60 (1973), pp. 255-265.

[89] P. Kaminski, A. Bryson, and S. Schmidt, "Discrete square root filtering- a survey of current techniques," *IEEE Trans. on Automatic and Control ,* No.16, pp. 727–736, 1971.

[90] D. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Atomatica ,* No.4, pp. 73–92, 1966.

[91] T. Mckelvey, H. Akcay, and L. Ljung, "Subspace-based identification of infinte-dimensional multivariable systems from frequency-response data," *Atomatica ,* No.32, pp. 885–903, 1996.

[92] T. Mckelvey, H. Akcay, and L. Ljung, "Subspace-based multivariable system identification from frequency-response data," *IEEE Trans. on Automatic and Control ,* Vol.41, pp. 960–976, 1996.

[93] P. Park and T. Kailath, "New square-root smoothing algorithms," *IEEE Trans. on Automatic Control ,* Vol.41, pp. 727–732, 1996.

[94] T. Sadiki and D. T. M. Slock, "Bayesian Adaptive Filtering: Principals and Practical Approaches," *Proc. Eusipco ,* 16-17 September. 2004.

[95] R. H. Shumway and D. S. Stoffer, "An Approach to Time Series Smoothing and Forecasting using The EM Algorithm," *Journal of Time Series Analysis ,* Vol. 3, No. 4,1982.

[96] Jerry M. Mendel, "Lessons In Estimation Thory For Signal Prossing, Communications, And Control,"Prentice Hall Signal Prossing Series, 1995.

[97] Jansson M., Goransson, B. and Ottersten, B., "A subspace method for direction of arrival estimation of uncorrelated emitter signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing ,* Vol. 47, No. 4, pp. 945–956, Apr. 1999.

[98] D. M. Titterington, "Recursive Parameter estimation using Incomplete Data," *J. R. Statist. Soc. ,* Vol. 46, No. 2, pp.257-267, 1984.

[99] C. F. Jeef Wu, "On The Convergence Properties of the EM Algorithm," *The Annals of Statistics ,* Vol. 11, No. 1, pp.95-103, 1983.

[100] Y. Zhao, "An EM Algorithm for Linear Distortion Channel Estimation Based on Observations for a Mixture of Gaussian Sources," *IEEE Trans. on Speech and Audio Processing,* Vol. 7, No. 4, Jul. 1999.

[101] Dempster A. P., Laird N. M. and Rubin D. B. "Maximum Likelihood from incomplete data via EM algorithm," *Journal Roy. Stat. Soc. Ser.,* Vol.39, No. 1, pp. 1–38 1977.

[102] Moon T. K., "The Expectation-Maximization algorithm," *IEEE Signal Proc. Mag.* , Vol.13, , pp. 47–60 1996.

[103] L. ljung, "Asymptotic behaviour of the extended Kalman filter as a parameter estimator for linear systems.," *IEEE Trans. on Automatic Control ,* Vol.24, , pp. 36–50 1979.

[104] T. Anderson and J. Taylor, "Strong consistency of least-squares estimates in dynamic models," *Annals of Statistics,* 7 (1979), pp. 484-489.

[105] T. Anderson, "The Statistical Analysis of Time Series," John Wiley and Sons, 1971.

[106] J. Rissanen and P. Caines, "Strong consistency of maximum likelihood estimators for ARMA process," *Annals of Statistics,* 7 (1979), pp. 297-315.

[107] W. Favoreel, "Subspace Methods for Identification and Control of Linear and Bilinear Systems," *PhD thesis,* Departement Elektrotechniek, Katholieke Universiteit Leuven, 1999.

[108] W. Favoreel, B. Moor, and P. van Overschee, "Subspace identification of bilinear systems subject to white inputs," *IEEE Transactions on Automatic Control,* 44 (1999), pp. 1157-1165.

[109] V. Verdult and M. Verhaegen, "Subspace identification of multivariable linear parameter-varying systems," *Automatica,* 38 (2002), pp. 805-814.

[110] T. Gustafsson,"System identification using subspace-based instrumental variable methods," *in Proc. of the 11th IFAC Symposium on System Identification,* SYSID 97, vol. 3, Kitakyushu, Japan, July 8-11 1997, pp. 1119-1124.

[111] K. Peternell, "Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms," *PhD thesis,* TU Wien, 1995.

[112] D. Bauer and L. Ljung, "Some facts about the choice of weighting matrices in Larimore type of subspace algorithms,"*Automatica,* 38 (2002), pp. 763-773.

[113] D. Bauer, M. Deistler, and W. Scherrer, "Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs," *Automatica,* 35 (1999), pp. 1243-1254.

[114] M. Deistler, K. Peternell, and W. Scherrer, " Consistency and relative efficiency of subspace methods," *Automatica,* 31 (1995), pp. 18651875.

[115] Jiang, C. J., "The use of mixture models to detect effects of major genes on quantitative characteristics in a plant-breeding experiment," *Genetics,* 136(1): 383-94, 1994.

[116] Redner, R. A. and H. F. Walker, "Mixture densities, maximum-likelihood estimation and the EM algorithm," *SIAM Rev.,* 26(2): 195-237, 1984.

[117] Schmee, J., and G. J. Hahn, "Simple method for regression analysis with gensored data" *Technometrics,* 21 (4): 417-32, 1979.

[118] Little, R. J. A., and D. B. Rubin, "On jointly estimating parameters and missing data by maximizing the complete-data likelihood," *Am. Statistn.,* 37(3): 218-200, 1983.

[119] Luenberger, David G.," Optimization by vector space methods," *Wiley,* New York, 1969.

[120] Rabiner, L. R., "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE,* 77(2): 257-86, Feb., 1989.

[121] Segal, M., and E. Weinstein, "Parameter estimation of continuous dynamical linear systems given discrete time observations," *Proc. IEEE,* 75(5): 727-29, 1987.

[122] Roy, R., A. Paulraj, and T. Kailath, "A subspace rotation approach to estimation of parameters of cisoids in noise,"*IEEE Trans. Acoust., Speech, S.P.,* 34(5): 1340-42, Oct., 1986.

[123] Ziskind, I., and M. Wax, "Maximum likelihood localization of narrow-band autoregressive sources via EM algorithm," *IEEE Trans. S.P.,* 41(8): 2719-24, 1993.

[124] Lagendijk, R. L., J. Biemond, and D. E. Boekee, "Identification and restoration of noisy blurred images using the expectation-maximization algorithm" *IEEE Trans. Acoust., Speech, S.P.,* 38(7): 1180-91, 1990.

[125] Ansari, A., and R. Viswanathan, " Application of EM algorithm to the detection of direct sequence signal in pulsed noise jamming," *IEEE Trans. comm., ,* 41(8): 1151-54, 1993.

[126] Feder, M., "Parameter estimation and extraction of helicopter signals observed with a wide-band interference," *IEEE Trans. S.P., ,* 41(1): 232-44, 1993.

[127] Byrne, W., "Alternating Minimization and boltzman machine learning," *IEEE Trans. Neural Network,* 3(4): 612-20, 1992.

[128] Jordan, M. I., and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural comp.,* 6(2):181-214, 1994.

[129] V, M. I., and D. R. Fuhrmann, "Maximum likelihood narrow-band direction finding and the EM algorithm,"*IEEE Trans. Acoust., Speech, S.P.,* 38(9): 1560-77, 1990.

[130] Vaseghi, S. V., and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *IEEE Trans. proc.,* 137(1):38-46, 1990.

[131] Weinstein, E., et al., "Iterative and sequential algorithms for multisensor signal enhancement,"*IEEE Trans. Sig. Proc.,* 42(4): 846-59, 1994.

[132] Gerghiades, C. N., and D. L. Snyder, "The Em algorithm for symbol unsynchronized sequence detection,"*IEEE Trans. comm.,* 39(1): 54-61, 1991.

[133] Antoniadis, N., and A. O. Hero, "Time delay estimation for filtered poisson processes using an EM-type algorithm,"*IEEE Trans. Sig. Proc.,* 42(8): 2112-23, 1994.

[134] Segal, M., and E. Weinstein," The cascade EM algorithm,"*IEEE Trans. Proc.,* 76(10): 1388-90, 1988.

[135] Blahut, R. E., "Computation of channel capacity and rate-distortion functions," *IEEE Trans. info. Theory,* 18(4): 460-73, Jul., 1972.

[136] Csiszar, I., and G. Tusnday, "Information geometry and alternating minimization procedures," *Statistics and Decisions, Supplement Issue,* 205-337,1984.

[137] O. M. Machhi and N. J. Bershad, "Adaptive Recovery of a Chirped Sinusoid in Noise: I. Performance of the RLS Algorithm," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.39, pp. 583–594, March 1991.

[138] S. Haykin, A. H. Sayed, J. R. Zeidler, P. Yee, P. C. Wei, "Adaptive Tracking of Linear Time-Variant Systems by Extended RLS Algorithms," *IEEE Trans. Signal Processing*, Vol.45, pp. 1118-1128, May 1997.

[139] K. Maouche and D. T. M. Slock, "Performance Analysis and FTF Version of the Generalized Sliding Window Recursive Least-Squares (GSWRLS) Algorithm," *In Proc. Asilomar Conference on Signals, Systems and Computers*, Vol.1, pp. 685–689, Pacific Grove, USA, Nov. 1995.

[140] K. Maouche, "Algorithmes des Moindres Carrés Récursifs Doublement Rapides : Application à l'Identification de Réponses Impulsionnelles Longues," *PhD Thesis*, Ecole Nationale Supérieure des Télécommunications, Paris, March 1996.

[141] M. Montazeri, "Une Famille d'Algorithmes Adaptatifs Comprenant les Algorithmes NLMS et RLS. Application à l'Annulation d'Echo Acoustique," *PhD Thesis*, Université de Paris-Sud, U.F.R Scientifique d'ORSAY, Sept. 1994.

[142] E. Eleftheriou and D. Falconer, "Tracking Properties and Steady State Performance of RLS Adaptive Filter Algorithms," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol.34, No.5, pp. 1097–1110, Oct. 1986.

[143] M. Niedzwiecki, "First-Order Tracking Properties of Weighted Least Squares Estimators," *IEEE Trans. Automatic Control*, Vol.33, No.1, pp. 94–96, Jan. 1988.

[144] M. Niedzwiecki, "Identification of Time-Varying Processes," *Wiley & Sons*, 2000.

[145] M. Lenardi and D. Slock, "Estimation of Time-Varying Wireless Channels and Application to the UMTS W-CDMA FDD Downlink,"

[146] Y. A. Rozanov, "Stationary Random Processes," *Holden-Day,* San Francisco, 1967.

[147] S. Tayeb, T. Mahdi and S. T. M. Dirk,"Window optimization issues in recursive least-squares adaptive filtering and tracking" *Asilomar 2004, 38th IEEE Annual Asilomar Conference on Signals, Systems and Computers,* November 7-10, 2004, Pacific Grove, USA.

[148] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME Series D: J. Basic Eng.,* 82 (1960), pp. 35 45.

[149] E. Kamen and J. Su, "Introduction to Optimal Estimation," *Springer-Verlag London Limited,* 1999.

[150] E. Hannan and M. Deistler, "The Statistical Theory of Linear Systems," *John Wiley and Sons,* New York, 1988.

[151] A. H. Jazwinski, "Stochastic Processes and Filtering Theory," *Academic Press,* 1970.

[152] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," *McGraw Hill,* London, 3 ed., 1991.

[153] B. Anderson and J. Moore, "Optimal Filtering," *Prentice Hall,* 1979.

[154] J. Doob, "Stochastic Processes," *John Wiley and Sons,* London, 1953.